



## OPEN Validation of reading as a predictor of mild cognitive impairment

Vida Groznik<sup>1,2,✉</sup>, Martin Možina<sup>2</sup>, Timotej Lazar<sup>1</sup>, Dejan Georgiev<sup>2,3</sup>, Aleš Semeja<sup>1</sup> & Aleksander Sadikov<sup>1,2</sup>

Mild cognitive impairment (MCI) is a neurocognitive disorder that precedes Alzheimer's disease, but also other types of dementia. The use of reading tasks, when paired with eye-tracking technology, has been suggested as an effective biomarker for identifying MCI and distinguishing it from healthy individuals. The objective of this study was twofold: (1) to explore the disparities in eye movements during reading between individuals with MCI and healthy controls and train a predictive model to detect MCI, and (2) to validate these findings on a large independent dataset. We developed features for a model designed to automatically detect cognitive impairment based on the data of 115 subjects; 62 cognitively impaired and 53 healthy controls. Each subject was subjected to a neurological evaluation, a thorough psychological analysis, and completed a brief reading exercise while their eye movements were monitored using an eye-tracker. Their eye movements were characterised by patterns of saccades and fixations and were analysed across both groups. Several characteristics showed very high statistical significance, indicating differences in gaze behaviour between the groups. These characteristics were then employed to develop a machine learning model that differentiates cognitively impaired individuals from healthy controls. For the validation purposes, we ran a separate study with 99 new subjects using the same experimental design. The model reached about 75% AUROC. These results confirm that reading tasks can serve as a basis for early detection of MCI; however, complementary eye-tracking tasks are needed to further increase the detection accuracy.

**Keywords** Eye-tracking, Machine learning, Mild cognitive impairment (MCI), Validation, Reading characteristics

Mild cognitive impairment (MCI) is a neurocognitive condition characterised by cognitive deficits that exceed what might be expected for a person's age and educational background, but is not severe enough to disrupt instrumental daily activities. This impairment exists without other identifiable causes, such as neurological, psychiatric, systemic disorders, metabolic dysfunction, or medication effects<sup>1</sup>. MCI was first used to describe the transition between normal cognition and dementia in Alzheimer's disease (AD)<sup>2</sup> but precedes other types of dementia as well<sup>1</sup>. In clinical cohorts, the annual conversion rates from MCI to AD range between 10-15%<sup>2</sup>.

Reading is a sophisticated cognitive activity that relies on numerous mental abilities, including attention, visual and auditory processing, as well as working, short- and long-term memory, and sensory integration<sup>3</sup>. Several studies have examined reading ability in MCI in combination with eye-tracking. In one of the first studies to examine eye movement behaviour in patients with cognitive decline, the eye movements of 14 patients with mild to moderate AD were recorded and compared to the eye movements of 14 healthy controls (HC) while they read a predefined text<sup>4</sup>. Compared to HC, AD patients did not differ in saccade duration; they had significantly longer fixation time, more forward saccades per line of text, and more saccadic regressions. Fernández et al.<sup>5</sup> reported patients with early AD who had a lower number of words with only one fixation, a higher total number of fixations in the first and second pass, a higher total number of saccadic regressions, and a higher number of skipped words. They also had a smaller size of outgoing saccades and, at the same time, a longer fixation duration. In another study by the same group<sup>6</sup>, gaze duration changed in HC as a function of sentence predictability (the higher the predictability, the shorter the gaze duration). AD patients showed an increase in gaze duration compared to HC, both for sentences with low and high predictability. The information from these studies was later used to build a machine learning based model<sup>7</sup> that could identify early AD with up to 89.78% accuracy. A further extension of this work<sup>8</sup> on a different sample of patients with MCI and HC used the same basic descriptors as Biondi et al. (2017)<sup>7</sup>, but modified by using frequency information (how often a word appears in the text). Similar accuracy of 86% was obtained in distinguishing MCI from HC, confirming

<sup>1</sup>NEUS Diagnostics, d.o.o., Ljubljana, Slovenia. <sup>2</sup>Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia. <sup>3</sup>Department of Neurology, University Medical Centre Ljubljana, Ljubljana, Slovenia. ✉email: vida.groznik@fri.uni-lj.si

the feasibility of using machine learning approaches to distinguish individuals with (mild) cognitive decline from healthy controls. These results were later confirmed by another study by the same group<sup>9</sup> where their best multimodal language classifier and the neuropsychological classifier reached the AUROC = 0.90 and CA = 0.84. Reviews<sup>10,11</sup> on this topic also suggest that eye movements can be used to discriminate between MCI and HC. Furthermore, this approach can be used to predict the transition from MCI to AD, as eye movements are sensitive to subtle changes in memory, visual, and executive functions<sup>10</sup>. Taken together, the available data suggest that reading combined with eye-tracking is a promising tool to discriminate between subjects with and without cognitive impairment (CI).

Most of the studies mentioned tested the use of this approach in patients with early AD; only a few studies included patients with MCI. In addition, the number of patients (MCI/AD) and HC in the mentioned studies is rather small (ranging from 14–40). In addition, only three studies<sup>7–9</sup> used machine learning approaches to discriminate between patients and HC. One of the prerequisites for the successful use of machine learning algorithms is a large sample size. Furthermore, it is not clear which basic descriptors derived from eye-tracking or combined with neuropsychological and/or clinical data are optimal to discriminate between MCI and HC.

Individuals with AD have been suggested to experience altered reading abilities due to primary oculomotor abnormalities such as instability in fixation<sup>4,12</sup>. Those with AD progressively show impairments in both inhibitory control and correction of eye movement errors<sup>12</sup>. In particular, the capacity to voluntarily shift the gaze away from a conspicuous stimulus in the antisaccade task (AST) is compromised in AD. Moreover, the frequency of errors in the AST is correlated with the severity of AD<sup>12</sup>. Eye movement deficits may appear early in the disease's progression, well before cognitive impairments become apparent through standard neuropsychological assessments<sup>13</sup>. Consequently, the question arises as to whether eye-movement impairment can be detected in individuals with MCI. Furthermore, eye movement disorders have been described in other neurodegenerative types of dementia, including frontotemporal dementia (FTD)<sup>14</sup> and Lewy body dementia (LBD)<sup>14,15</sup>. To the best of our knowledge, no studies have been performed so far on eye-tracking during reading in MCI preceding either FTD or LBD.

Identifying MCI using inexpensive and simple tools is thus essential, since invasive and/or expensive biomarkers may not be widely available, especially for population-level usage. Eye-tracking is a promising method that fulfils these criteria. Combined with reading, a task that is heavily dependent on impaired cognitive functions in the early stages of the disease, eye-tracking could be the ideal approach for convenient detection of cognitive decline.

In this study, we investigated the differences in eye movement during reading between people with various levels of cognitive impairment and healthy controls. Based on the observed differences, we then designed the features to use in a predictive model for automatic detection of MCI. Finally, we conducted a validation study to test the predictive model on a completely independent dataset.

The paper is structured as follows. First we report on study participants and their characteristics in Section 2 and describe the machine learning setup in Section 3. Section 4 presents the results of statistical comparisons and initial machine learning, while Section 5 presents the results of the independent follow-up validation study. The final two sections discuss our results and conclude the paper.

A preliminary version of this work has been reported in a paper<sup>16</sup> accepted for presentation at the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) 2021.

## Study subjects

The data used in this analysis was collected in two separate studies. The first study was performed to explore reading as a biomarker and to fit a predictive model. The follow-up study collected data from new participants, used to validate the proposed model. The experimental design was nearly identical for both studies. Based on user and expert feedback from the first study, some eye-tracking tasks were updated for the second one; however, the reading task investigated in this paper remained unchanged.

## Enrollment criteria

In the initial study, subjects were recruited from nursing homes for elderly. Recruitment occurred either through physician referrals — when individuals expressed concern about their potential cognitive issues (subjective cognitive impairment) — or through volunteers who learned about the study from other participants or directly from the researchers. The follow-up study was conducted at a private medical centre in Slovenia, using the same recruitment methods as the initial study.

The following inclusion criteria were observed when recruiting participants:

- subjects with CI older than 40 years;
- subjects with no CI older than 40 years.

At the same time, the following exclusion criteria were used for recruitment:

- subjects with uncorrected visual impairments;
- subjects with concomitant neurological diseases;
- subjects with psychiatric diseases, including those with GDS-15 (Geriatric Depression Scale – 15 questions<sup>17</sup>) > 10;
- drug or alcohol abuse;
- subjects who declined to participate in the study or were unable to perform all the tests.

The data was collected under a clinical study approved by the National Medical Ethics Committee of the Republic of Slovenia (approval numbers: 0120-400/2015-5 dated April 2, 2016; 0120-400/2015/9 dated May 22, 2018; and 0120-400/2015/17 dated June 26, 2019; the Committee also issued a notice of notifications and minor changes no. 0120-400/2015/20, dated September 17, 2019). The clinical studies were conducted in accordance with a clinical protocol, good clinical practice and applicable national regulations to ensure that the rights, safety and welfare of the subjects were protected in accordance with the ethical principles enshrined in the Declaration of Helsinki.

The purpose and procedures of the study were explained to the subjects verbally and in writing prior to inclusion in the study. Based on this information, they were free to choose to participate in the study without any pressure or conditions. All subjects took part in the study voluntarily and were informed of the possibility of terminating the study or withdrawing their consent at any time.

### Subjects' characteristics

This subsection describes the characteristics of subjects in the first study, used for exploration and modelling. The characteristics of subjects in the validation study are given alongside the validation results in Section 5 for ease of understanding.

We analysed data from 115 subjects who were *consecutively* recruited for the study. The youngest subject was 43 years old, and the oldest was 94, with a median age of 68 years. The dataset comprised 85 female and 30 male subjects.

Diagnoses were made using the DSM-V criteria<sup>18</sup>. Based on neurological and psychological assessments (described below), each participant was classified into one of the following categories: healthy (no cognitive decline), borderline (some cognitive decline detected but not sufficient for an MCI diagnosis), MCI (mild cognitive impairment), or possible Alzheimer's disease (AD).

Subjects with MCI needed to have at least one out of six cognitive domains affected (memory and learning, complex attention, executive function, language, perceptual-motor, or social cognition) with no impairments in basic activities of daily living and no dependence on others in any instrumental activities of daily living. Possible AD was diagnosed as impairment in at least one cognitive domain, but with impairments in basic activities of daily living and dependence on others in the instrumental activities of daily living. Subjects having some impairment in one cognitive domain, but not enough to diagnose MCI were designated as borderline.

A total of 53 subjects were diagnosed as healthy, 32 as borderline, 19 as having MCI, and 11 as possible AD. For the purpose of machine learning, we defined a binary classification problem with two groups: (1) healthy controls (HC), and (2) cognitively impaired (denoted as CI), which includes borderline, MCI, and subjects with possible AD. Table 1 provides further details on the age and gender distribution for each diagnosis/group.

### Subjects assessment

Each subject underwent an assessment lasting approximately two hours, which was conducted in three steps:

#### 1. Neurological examination

This step involved evaluating the participants' cognitive condition and, when necessary, their motor and non-motor functions. Demographic information was gathered through a questionnaire that included details about family and medical history, current treatments, potential head injuries, and other relevant factors.

	Healthy	Borderline	MCI	Possible AD	CI
<i>N</i>	53	32	19	11	62
<i>Gender</i>					
Female	40	24	12	9	45
Male	13	8	7	2	17
<i>Age</i>					
Median	63	68.5	72	83	72
Range	48–83	60–87	43–91	72–94	43–94
<i>MMSE</i>					
Mean	28.98	28.38	26.79	23.64	27.05
SD	1.03	1.21	1.65	2.46	2.36
<i>ACE-R</i>					
Mean	94.26	87.41	81.53	62.36	81.16
SD	2.91	4.21	6.83	7.98	10.84
<i>GDS-15</i>					
Mean	1.25	1.56	2.16	3.18	2.03
SD	1.93	1.81	2.54	3.22	2.37

**Table 1.** Gender, age, and cognitive scores distribution per diagnosis/group.

## 2. Psychological assessment

The psychological assessment involved evaluating higher cognitive capabilities, including executive functions and memory by a certified clinical psychologist. To assess cognitive abilities, the following standardised tests were employed: ACE-R (Addenbrooke Cognitive Examination-Revised)<sup>19</sup>, FAB (Frontal Assessment Battery)<sup>20</sup>, CTMT (Comprehensive Trail Making Test)<sup>21</sup>, and GDS-15 (Geriatric Depression Scale – 15 questions)<sup>17</sup>.

Subjects with ACE-R score less than 86 (for subjects aged 50–59 years), or less than 85 (for subjects aged 60–69 years), or less than 84 (for subjects aged 70–79 years) and/or subjects with Mini Mental Score Examination (MMSE, part of ACE-R) score less than 26 were candidates for placement into either borderline, MCI, and possible AD group. In addition to participants' age, the education level was also taken into account in accordance with the used psychological tests. In cutoffs, we followed the established clinical practice, e.g. ACE-R<sup>19</sup>, MMSE<sup>22,23</sup>. The final decision about the group placement – also for subjects flagged as healthy by above mentioned tests – was taken by a neurologist when presented with complete information after the clinical examination of the patient and after observing the psychological report. More details about the cognitive scores distribution per diagnosis/group is given in Table 1.

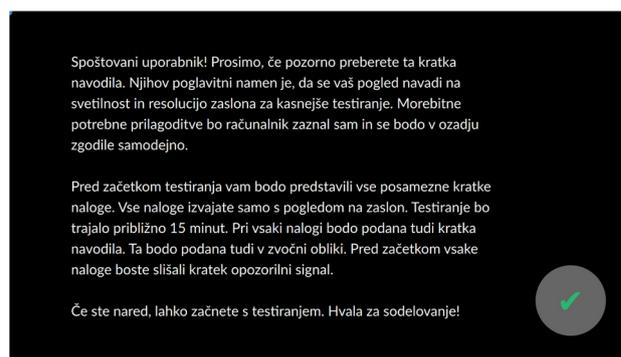
## 3. Eye-tracking test battery

The eye-tracking test battery was conducted using a 90 Hz Tobii 4C eye-tracker, specialised software developed by NEUS Diagnostics, d.o.o., and associated computer equipment — including a laptop for the examiner and a 23.6-inch display with a 1920×1080 pixel resolution for the participants. A trained technician administered the tests. Participants sat approximately 70 cm from the display and had no direct contact with the equipment, interacting with the test battery solely through their gaze. The NEUS software presented visual stimuli on the screen and guided participants through the entire test battery. It recorded their eye movements and stored the data in a database for later analysis. All data were anonymised, and access to the database was restricted to researchers involved in the study.

After the initial 5-point eye-tracker calibration, subjects were asked to read the text displayed on the screen in a large font. This task was presented as another part of the calibration process before the actual test, in order to reduce the stress subjects might feel due to testing – the reasoning is that they might perform the reading task more naturally. The text was arranged over 11 lines in three paragraphs, occupying most of the display area. After reading the text, subjects were instructed to explicitly confirm they have finished reading by directing their gaze at the button in the bottom right corner of the screen (see Figure 1). The reading task on average took from less than half a minute for healthy subjects to just over a minute for patients with possible AD.

## Machine learning setup

The main machine learning objective was to try to differentiate between CI subjects and HC based on their gaze behaviour during reading. We have thus decided on a binary class, CI versus HC. As previously described, there were 62 subjects in the CI group and 53 in the HC group. We used only the nine reading features described below without any additional data to learn from. The reason for this decision is that we wanted to investigate how well a fast, stand-alone reading test can work on its own without collecting any additional data.



**Fig. 1.** The short reading task (in Slovenian), comprising brief instructions and a confirmation button in the bottom right corner. The translation is as follows: “Dear user! Please read these brief instructions carefully. Their main purpose is to get your eyes used to the luminance and resolution of the computer screen for later testing. Any necessary adjustments will be detected by the computer itself and will happen automatically in the background. Before the start of the testing, we will present each of the short tasks to you. All tasks are performed by looking at the screen only. The test will take approximately 15 minutes. Brief instructions will be given for each task. These will also be given in audio form. You will hear a short warning signal before the start of each task. If you are ready, you can start with the test. Thank you for your participation!”

## Eye-tracking reading features

The raw eye-tracking data are an event series of  $(x, y)$  screen coordinates and the accompanying timestamp  $t$ , sampled at 90 Hz. We transformed the raw data into higher-level features using domain knowledge to (1) better facilitate the machine learning process, and to (2) gain human comprehensible insights into differences between observed groups.

### Pre-processing

First, we discarded events with invalid coordinates — instances where the eye-tracker failed to detect the eye due to blinking, the user looking away from the screen, or similar issues. We also excluded the initial 5% and the final 10% of each recording since gaze behaviour can be erratic at the start and end of a recording session. These two thresholds were set based on the data from the pilot study (not reported here).

After cleaning the eye-tracking data, we extracted fixations and saccades. A fixation was defined as maintenance of the gaze on a single point, while a saccade was defined as a rapid, simultaneous movement of both eyes between two fixation points. We defined fixations as sequences of consecutive samples lasting at least 48 milliseconds with inter-sample distances no greater than 100 pixels (27.2 mm). These parameters were empirically determined to suit our sampling frequency and the relative size of the text displayed on the screen. Each pair of consecutive fixations identified in this manner define one saccade.

### Definitions and meaning of reading features

Using the extracted fixations and saccades, we defined several reading features: the distances of forward and backward saccades, the variability in backward saccade distances, the speeds of forward and backward saccades, fixation durations, the variation in fixation durations, the ratio of forward to backward saccades, and the number of fixations per second. The goal of selecting these features was to capture as many aspects of the reader's behaviour as possible while maintaining a reasonable number of variables. Detailed descriptions of these features are given in<sup>16</sup> and are also provided below for easier understanding of this paper.

Direction of a saccade can be either forward (right) or backward (left). It is detected from the screen positions of the fixations before and after the saccade in question. The forward saccade distance (*fdist*) and backward saccade distance (*bdist*) features are defined as the median distance of all detected saccades in the given direction. These features correlate with the speed of reading, but also contain additional information, e.g., *bdist* is closely associated with starting a new line of text or with re-reading some parts of the text.

The amount of variation in backward saccade distances (*bdist.stdev*) is the standard deviation of the distances of all detected backward saccades. As already mentioned, backward saccades are closely associated with starting a new line of text and thus tend to be long. However, variation in these distances can signal that the user is re-reading some words, or is lost in the text, etc.

The forward saccade speed (*fspeed*) and backward saccade speed (*bspeed*) are defined as the median speed of all detected saccades in the given direction. While speed and distance correlate with each other, they are not the same, so these features can offer additional information into the reading behaviour.

The duration of fixations (*fdur*) is defined as the median duration of all detected fixations, and the amount of variation in fixation duration (*fdur.std*) as the standard deviation of the durations of all detected fixations. Both of these features can be associated with erratic or confused behaviour of the reader.

The ratio between forward and backward saccades (*fsVbs*) is computed as the number of detected forward saccades divided by the number of detected backward saccades.

The number of fixations per second (*fixt*) is the time-normalised number of detected fixations. This feature is a more robust version of reading speed (or time spent during reading).

## Algorithms

There was no additional pre-processing of the data beyond the feature construction process described in the previous section. The only exception was excluding one case from the HC group that was an extreme outlier due to missing data despite multiple recalibrations.

We have explored four machine learning algorithms: logistic regression (LR), naïve Bayes (NB), stochastic gradient boosting (XGBoost algorithm; XGB), and random forest (RF). Each algorithm's parameters were optimised with internal cross validation. The complete machine learning pipeline was implemented using the scikit-learn v0.23.1 library.

The algorithms were evaluated using the following measures: classification accuracy (CA), Brier score (Brier)<sup>24</sup>, area under the receiver-operating-characteristic curve (AUC), sensitivity, and specificity. We used 10-times repeated 10-fold cross validation. We also used the majority classifier (dummy; always classifying into the majority class) for comparison.

## Results

In this section we first explore how various feature values are distributed among groups, both as a binary problem (HC versus CI) and with increasing level of cognitive impairment (from HC to possible AD). Next, we present the machine learning results of distinguishing between HC and CI. The results in this section are based on the initial study involving 115 subjects. Data from the second study, involving 99 new subjects, was intentionally withheld exclusively for validation purposes and is presented in its own section later.

### Statistical distributions of reading features with increasing level of cognitive impairment

We were interested in how the reading features are distributed along all four groups of participants, from healthy controls to subjects with possible AD. Since most feature values were not normally distributed, we used the nonparametric Mann-Whitney U-test to make these comparisons. We report the results of the omnibus test

comparing all four groups. The significance level for all comparisons in this study was set at 0.05 and all tests were two-tailed.

From the graphs in Figure 2 and associated p-values we can see that all the features except *bdist.std* exhibit different distributions between groups. It can further be seen that the groups are ordered with respect to the level of cognitive impairment, progressing from HC to borderline and MCI groups and finally to subjects with possible AD. This result, although expected, is nevertheless interesting and important to confirm in practice.

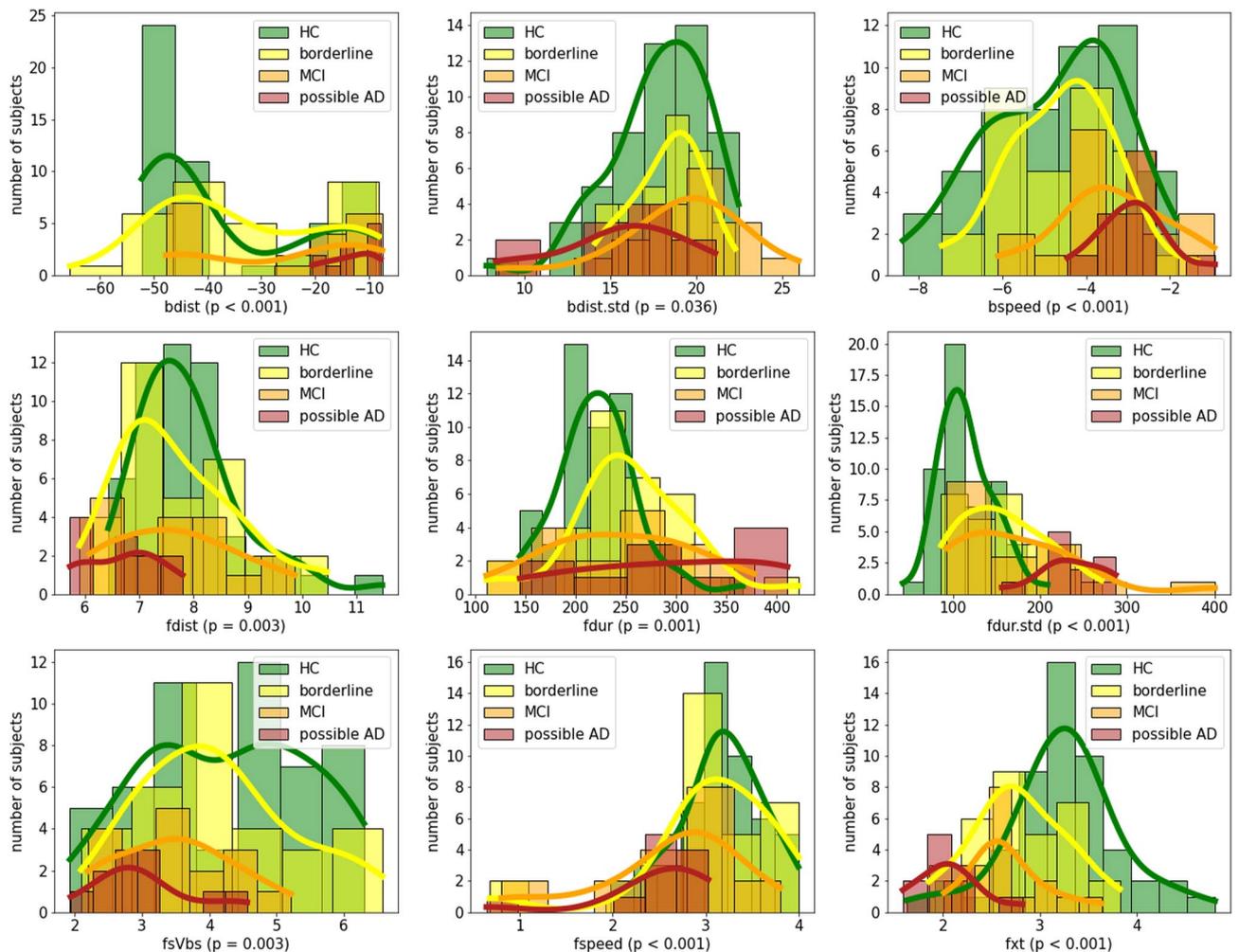
### Statistical distributions of reading features between HC and CI

While it is of clinical interest to observe how reading features are distributed with increasing level of cognitive impairment, it is the binary problem of detecting *any* level of cognitive impairment that is of primary interest when designing a decision support system for early detection of the disease. Therefore we also compared the distributions of the feature values between CI and HC groups. As before, we used the nonparametric Mann-Whitney U-test to test for differences. The results of these comparisons are reported in Table 2, along with some parameters of the feature value distributions.

Except for the feature *bdist.std*, all other metrics consistently indicate significant differences in eye movement behaviour between the two groups. Healthy controls (HC) tend to have a notably higher number of fixations per unit of time, shorter fixation durations, and a more uniform distribution of these durations. These three features are particularly distinct between the groups. Additionally, HCs make longer saccades in both forward and backward directions and perform them slightly faster. They also exhibit a higher ratio of forward to backward saccades.

### Machine learning results

The results from the 10-fold cross-validation, presented in Table 3, indicate that logistic regression is the top-performing algorithm overall. Naïve Bayes, however, slightly surpasses it in the AUC metric. Furthermore, the two algorithms seem to trade sensitivity and specificity between them. Specifically, logistic regression achieves a classification accuracy of 73.1%, an AUC of 79.7%, and a Brier score of 0.189. The confusion matrix for logistic



**Fig. 2.** Distributions of reading features over all four groups alongside with their associated p-values.

Feature	p-value	Group	Median	St. dev.	Min	Max
<i>bdist</i> [% of screen]	0.002	HC	-45.91	15.86	-52.23	-8.75
		CI	-20.51	16.21	-65.59	-7.48
<i>bdist.std</i> [% of screen]	0.512	HC	18.15	3.01	7.79	22.45
		CI	18.44	3.19	8.34	25.95
<i>bspeed</i> [mm/ms]	0.023	HC	-1.17	0.46	-2.28	-0.10
		CI	-1.05	0.39	-2.02	-0.25
<i>fdist</i> [% of screen]	0.017	HC	7.74	1.01	6.43	11.48
		CI	7.32	1.10	5.72	10.46
<i>fdur</i> [ms]	< 0.001	HC	216.51	41.87	127.69	366.41
		CI	255.37	70.92	111.03	421.92
<i>fdur.std</i> [ms]	< 0.001	HC	111.48	33.91	42.51	220.09
		CI	165.96	62.27	86.38	400.54
<i>fsVbs</i> [unitless]	0.026	HC	4.43	1.26	1.93	6.30
		CI	3.62	1.07	1.94	6.57
<i>fspeed</i> [mm/ms]	0.008	HC	0.86	0.17	0.06	1.09
		CI	0.78	0.21	0.17	1.09
<i>fxt</i> [#fixations/s]	< 0.001	HC	3.28	0.67	0.71	4.80
		CI	2.58	0.51	1.59	3.82

**Table 2.** Distributions of eye-tracking feature values between CI and HC groups.

Measure	Majority	LR	NB	XGB	RF
CA	0.544 ± .128	<b>0.731 ± .122</b>	0.723 ± .133	0.693 ± .137	0.729 ± .123
Brier	0.252 ± .013	<b>0.189 ± .047</b>	0.231 ± .108	0.220 ± .067	0.194 ± .064
AUC	0.500 ± .000	0.797 ± .140	<b>0.799 ± .135</b>	0.739 ± .152	0.783 ± .146
Sensitivity	1.000 ± .000	<b>0.778 ± .169</b>	0.682 ± .201	0.695 ± .189	0.740 ± .174
Specificity	0.000 ± .000	0.675 ± .206	<b>0.773 ± .194</b>	0.699 ± .213	0.722 ± .202

**Table 3.** The performance estimation of classification algorithms.

	Predicted class				Predicted class		
		HC	CI			HC	CI
True class	HC	<b>36</b>	16	True class	HC	<b>37</b>	16
	CI	15	<b>47</b>		CI	15	<b>26</b>

**Table 4.** Confusion matrices for the best performing logistic regression classifier. (left: initial study; right: validation study).

regression, shown in Table 4, reveals the types of misclassifications it makes. As evident from the matrix, the algorithm errs equally in both classes, which is expected given the relatively balanced nature of the dataset.

### Independent validation

As already mentioned, we performed an independent clinical study to validate our findings. This validation study observed the same protocol as the original study.

The validation study was officially registered with the Agency for Medicinal Products and Medical Devices of the Republic of Slovenia (approval no. 340-4/2019-8 dated 27 August, 2019). As such, the outcome of the examination had to be defined for all subjects – meaning that the borderline assessment was not an option in this validation study. The participants were therefore diagnosed as either HC, MCI, or possible AD.

We have recruited 99 *consecutive* new subjects for the validation study. Five of the subjects had other diagnoses (e.g. anxiety or depression) and were not eligible for the analysis. This left us with 94 subjects.

The youngest subject was 46 and the oldest 86 years old. The median age of the subjects was 69 years. The data consists of 60 female and 34 male subjects. Based on neurological and psychological evaluation, the validation set included 53 HC, 29 subjects with MCI, and 12 subjects with dementia. The latter two groups were merged into a single group of subjects with CI (as before denoted as CI). The detailed age and gender distribution of the subjects based on their diagnoses/group is given in Table 5 along with the cognitive scores distribution.

	Healthy	MCI	Possible AD	CI
<i>N</i>	53	29	12	41
<i>Gender</i>				
Female	36	17	7	24
Male	17	12	5	17
<i>Age</i>				
Median	67	70	72.5	72
Range	46–77	60–86	56–81	56–86
<i>MMSE</i>				
Mean	29.00	27.62	25.17	26.90
SD	0.83	1.29	3.19	2.29
<i>ACE-R</i>				
Mean	92.57	85.76	78.83	83.73
SD	3.58	7.14	8.55	8.12
<i>GDS-15</i>				
Mean	1.55	2.66	3.17	2.80
SD	1.81	2.61	2.08	2.45

**Table 5.** Gender, age, and cognitive scores distribution per diagnosis/group for subjects in the validation dataset.

Measure	Majority	LR	NB	XGB	RF
CA	0.564	0.670	0.670	<b>0.691</b>	0.681
Brier	0.258	0.204	0.294	0.217	<b>0.196</b>
AUC	0.500	0.741	0.715	0.694	<b>0.752</b>
Sensitivity	1.000	<b>0.634</b>	0.561	0.512	0.561
Specificity	0.000	0.698	0.755	<b>0.831</b>	0.774

**Table 6.** The performance of the classification algorithms on validation study data.

The model to differentiate between subjects with HC and CI, regardless of the level of impairment, which was learned from the data of 115 subjects and evaluated using cross-validation in the previous section, was not changed. The new data was used solely for the evaluation of its performance.

The results of the new data set are given in Table 6. The performance of all algorithms has decreased somewhat, with CA decreasing the most at about 5%. The LR and RF algorithms perform best and are quite similar in their performance.

The misclassifications remain as before approximately equal in both directions, HC to CI and vice versa (Table 4).

## Discussion

Our results show that gaze behaviour differs significantly between HC and subjects with at least some degree of CI. Compared to HC, CI subjects have significantly fewer fixations per time unit. Their fixations are on average longer, and the durations are not distributed as uniformly. This points to their reading behaviour being slower and especially more irregular in its appearance. Our findings are broadly in line with the earlier studies that investigated reading behaviour in individuals with AD or MCI<sup>4–6,25</sup>. Evidence from various studies and groups shows that eye-movement behaviour (including fixation duration, number of fixations, and the regularity of fixations) are robust indicators of CI.

Nevertheless, several methodological factors and language-related aspects can lead to variations in specific findings. First, the texts used in previous work have ranged from simple sentences to proverbs, potentially influencing the complexity and predictability of the reading materials. Second, while some studies focused on silent reading, others involved reading aloud, and the resulting eye-movement patterns can differ between these modes. Third, variations in how fixations, saccades, and regressions are defined or preprocessed (for example, thresholds for detecting fixations or noise removal procedures) inevitably affect the reported outcomes. Additionally, the languages used in the prior studies – such as Spanish in<sup>5,25</sup> vs. English in<sup>8</sup> – differ in their orthographic rules and word predictability patterns. Languages with more transparent orthographies or more inflectional complexity may shift the baseline for normal reading behaviour.

The graphs in Figure 2 show remarkably well that CI is really a continuum. Most of the constructed eye-tracking features increase or decrease smoothly from HC to borderline subjects, then MCI and finally to possible AD. We can see how the increasing level of impairment affects the gaze behaviour during reading: the fixations

become longer and more erratic, the number of fixations per time unit steadily decreases etc. The only exception is the *bdist.std* feature, which is not significant in a binary setting.

While this continuous gaze behaviour is in a way an expected result, it was still important to confirm it in practice as it further validates the use of eye-tracking to detect early stages of cognitive impairment. It also means that even if the impairment is not immediately detected, it will likely become more easily detectable as it worsens over time.

As is often the case, the practical utility of machine learning results depends on their intended application. Comparing our findings with a majority classifier makes it clear that the proposed reading features carry significant information for the early detection of MCI. Additionally, the near 80% AUC score achieved through cross-validation on the training dataset suggests that a clinical decision support system (CDSS) based on a brief reading test could be practically applicable. This system could serve either as a quick, standalone test with lower precision or be integrated into a more comprehensive and time-consuming neuropsychological test battery. The validation results on an independent dataset were about 5% worse, but within the expected range. The result of AUC near the 75% mark now points more towards the second option of using reading as part of a larger test battery including other neuropsychological tests.

Most of the previous studies presented in the introduction focused on using eye-tracking during reading in patients with early AD and only a few studies included patients with MCI. Whereas our main goal was to detect the signs of cognitive impairment as early as possible, so a great majority of our CI patients were diagnosed as either MCI or borderline. Additionally, the number of patients and HC in the mentioned studies is rather small (ranging from 14-40) and were targetly selected for the studies. On the other hand, our study included two independent cohorts. The first one with 115 subjects was used to build a diagnostic model to differentiate between HC and CI groups. The second cohort included 94 subjects and was used for validation and performance evaluation of the model built on the eye-tracking data from the first cohort. We believe a larger sample size (compared to other studies) is an important addition to the MCI detection research, since all accuracy measures are prone to vary, especially with smaller samples.

As mentioned before, reading could be a good addition to a larger test battery in combination with other neuropsychological tests. That said, the test battery should be as short as possible to not overburden the users while, at the same time, give relevant results. In this aspect a short reading task is most welcome. In previous studies the reading tasks consisted of reading a much longer text (75 sentences in<sup>5</sup>, 120 sentences in<sup>6</sup>, and<sup>26</sup>, 120 sentences and additional 64 proverbs in<sup>7</sup>, and two texts with 146 words per text in<sup>8,9</sup>. On the other hand, we used the reading task consisting of just 12 sentences making it suitable to be incorporated as one of the tasks into the envisioned test battery.

Reading is a universally practised human activity and is not an artificially designed test, which grants it substantial ecological validity in experimental settings. In our observations, participants — including those with dementia — did not struggle with reading a few paragraphs displayed on a computer screen. Conversely, some alternative tasks felt less natural and prompted numerous questions from the participants.

However, a drawback of utilising reading is the necessity to verify the language dependence or independence of the machine-learned models. While we believe that the features are general enough to be language-independent, this assumption requires confirmation. Moreover, this may not hold true for fundamentally different writing systems, such as Chinese or right-to-left scripts like Arabic. On the positive side, in this relatively large sample study we tested the use of eye-tracking based reading features in subjects with a spectrum of CI, including MCI and possible AD.

One could argue that the groups were not matched by age. However, the main objective of the study/model was to predict the eye movement abnormalities by the use of eye-tracking in subjects with cognitive decline (including borderline, MCI and AD) and without cognitive decline (healthy subjects) no matter the age. The whole sample consisted of participants randomly and consecutively recruited to participate in the study – and since cognitive decline is correlated with age, there are some differences in age between the subgroups. Stratifying for age would actually be quite difficult and would, on the other hand, introduce a bias (in distribution) away from the envisioned practical use (and we cannot satisfy these two competing objectives at the same time).

In certain types of dementia beyond Alzheimer's Disease (AD), such as Primary Progressive Aphasia (a subtype of Frontotemporal Dementia) language abilities are the first to deteriorate. Reading plays a significant role in language skills, alongside writing, naming, language production, and comprehension. In these disorders, the proposed reading features and a clinical decision support system based on them might perform even more effectively. Moreover, such a system could be utilised not only for the earliest possible detection of these conditions, but also to differentiate them from the altered reading abilities observed in other forms of dementia, such as AD, Lewy Body Dementia, and Vascular Dementia.

Despite the current lack of comparative studies, there are strong theoretical reasons to suspect that different neurodegenerative diseases might produce disease-specific signatures. AD, for instance, is frequently associated with pronounced memory deficits (particularly episodic memory), whereas Parkinson's Disease often involves distinct motor symptoms and executive dysfunction. Such clinical differences may well translate into different patterns in eye-tracking or linguistic measures (e.g., lexical retrieval, reading speed, error rates) if thoroughly investigated.

## Conclusions

The study confirmed the hypothesis that eye-movement behaviour during reading significantly differs between cognitively impaired individuals and healthy controls. Through proposed features it also gave further insight into the underlying nature of these differences.

Furthermore, the study confirmed that gaze behaviour during reading deteriorates with increasing level of CI. This further validates the eye-tracking approach proposed in this paper and hints that the detection will become increasingly more likely with increasing level of CI.

The machine learning experiments show that reading and the associated features proposed in this work are a promising approach to early detection of cognitive impairment using a simple, cheap, and noninvasive test. The somewhat lower results on the independent validation dataset indicate that the short task used in our study might not suffice as a stand-alone test. We have, however, shown that reading is a significant biomarker suitable as one of the fundamental components of a clinical decision support system.

### Data availability

The data that support the findings of this study are available from NEUS Diagnostics d.o.o. but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the corresponding author upon reasonable request and with permission of NEUS Diagnostics d.o.o..

Received: 22 October 2024; Accepted: 13 March 2025

Published online: 14 April 2025

### References

- Jongsiriyanong, S. & Limpawattana, P. "Mild cognitive impairment in clinical practice: a review article," *American Journal of Alzheimer's Disease & Other Dementias*. **33**(8), 500–507 (2018).
- Farias, S. T., Mungas, D., Reed, B. R., Harvey, D. & DeCarli, C. Progression of mild cognitive impairment to dementia in clinic-vic community-based cohorts. *Archives of neurology*. **66**(9), 1151–1157 (2009).
- Landi, N., Frost, S. J., Mencl, W. E., Sandak, R. & Pugh, K. R. Neurobiological bases of reading comprehension: Insights from neuroimaging studies of word-level and text-level processing in skilled and impaired readers. *Reading & Writing Quarterly*. **29**(2), 145–167 (2013).
- Lueck, K. L., Mendez, M. F. & Perryman, K. M. Eye movement abnormalities during reading in patients with Alzheimer disease. *Neuropsychiatry, Neuropsychology, & Behavioral Neurology*. **13**(2), 77–82 (2000).
- Fernández, G. et al. Eye movement alterations during reading in patients with early Alzheimer disease. *Investigative ophthalmology & visual science*. **54**(13), 8345–8352 (2013).
- Fernández, G. et al. Lack of contextual-word predictability during reading in patients with mild Alzheimer disease. *Neuropsychologia*. **62**, 143–151 (2014).
- Biondi, J., Fernandez, G., Castro, S. & Agamennoni, O. "Eye-movement behavior identification for AD diagnosis," arXiv preprint [arXiv:1702.00837](https://arxiv.org/abs/1702.00837), (2017).
- Fraser, K.C., Fors, K.L., Kokkinakis, D. & Nordlund, A. "An analysis of eye-movements during reading for the detection of mild cognitive impairment," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1016–1026, (2017).
- Fraser, K.C., Lundholm Fors, K., Eckerström, M., Öhman, F. & Kokkinakis, D. "Predicting MCI status from multimodal language data using cascaded classifiers," *Frontiers in aging neuroscience*. **11**, 205 (2019).
- Freitas Pereira, M.L., Camargo, M.v. Z.A., Aprahamian, I. & Forlenza, O.V. "Eye movement analysis and cognitive processing: Detecting indicators of conversion to Alzheimer's disease," *Neuropsychiatric disease and treatment*. **10**, 1273–1285 (2014).
- Beltrán, J., García-Vázquez, M.S., Benois-Pineau, J., Gutierrez-Robledo, L.M., & Dartigues, J.-F. "Computational techniques for eye movements analysis towards supporting early diagnosis of Alzheimer's disease: a review," *Computational and mathematical methods in medicine*. 2018, (2018).
- Wilcockson, T. D. et al. Abnormalities of saccadic eye movements in dementia due to Alzheimer's disease and mild cognitive impairment. *Aging* **11**(15), 5389–5398 (2019).
- Crawford, T. J. et al. Inhibitory control of saccadic eye movements and cognitive impairment in Alzheimer's disease. *Biological Psychiatry*. **57**(9), 1052–1060 (2005).
- Russell, L. L. et al. Eye movements in frontotemporal dementia: Abnormalities of fixation, saccades and anti-saccades. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*. **7**(1), e12218 (2021).
- Sekar, A., Panouillères, M.T., Kaski, D. "Detecting abnormal eye movements in patients with neurodegenerative diseases—current insights," *Eye and Brain*. 3–16, (2024).
- Grozniak, V., Možina, M., Lazar, T., Georgiev, D. & Sadikov, A. "Gaze behaviour during reading as a predictor of mild cognitive impairment," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)1–4*, IEEE, (2021).
- Conradsson, M. et al. Usefulness of the Geriatric Depression Scale 15-item version among very old people with and without cognitive impairment. *Aging & mental health*. **17**(5), 638–645 (2013).
- Association, A.P., Force, D.-T. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5™*. American Psychiatric Publishing, Inc., 5th ed., (2013).
- Mioshi, E., Dawson, K., Mitchell, J., Arnold, R. & Hodges, J. R. The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*. **21**(11), 1078–1085 (2006).
- Slachevsky, A. et al. Frontal assessment battery and differential diagnosis of frontotemporal dementia and Alzheimer disease. *Archives of neurology*. **61**(7), 1104–1107 (2004).
- Bowie, C. R. & Harvey, P. D. Administration and interpretation of the Trail Making Test. *Nature protocols*. **1**(5), 2277–2281 (2006).
- Kukull, W. et al. The mini-mental state examination score and the clinical diagnosis of dementia. *Journal of Clinical Epidemiology*. **47**(9), 1061–1067 (1994).
- Folstein, M.F., Folstein, S.E., McHugh, P.R. *MMSE Clinical Guide*. Psychological Assessment Resources, Inc., (2002).
- Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly weather review*. **78**(1), 1–3 (1950).
- Fernández, G., Castro, L. R., Schumacher, M. & Agamennoni, O. E. Diagnosis of mild Alzheimer disease through the analysis of eye movements during reading. *Journal of integrative neuroscience*. **14**(1), 121–133 (2015).
- Fernández, G. et al. Patients with mild Alzheimer's disease fail when using their working memory: evidence from the eye tracking technique. *Journal of Alzheimer's Disease*. **50**(3), 827–838 (2016).

### Acknowledgements

This research has received funding under project NEUS from the European Institute of Innovation and Technology (EIT) Health KIC. This body of the European Union receives support from the European Union's Horizon

2020 research and innovation programme. The research was also partially supported by Slovenian Research and Innovation Agency under the research programme Artificial intelligence and intelligent systems grant no. P2-0209. The authors would also like to thank all the neurologists, psychologists, technicians, and administrative support staff who were responsible for patient onboarding and data collection.

### Author contributions

VG, DG, ASe, and ASa wrote the main manuscript text; VG, MM, and ASa performed the data analysis and built machine learning models; TL programmed the software for data collection; VG, DG, and ASa prepared the study protocol; all authors reviewed the manuscript.

### Declarations

### Competing interests

VG, ASe, and ASa are co-owners of NEUS Diagnostics d.o.o., TL was employed by NEUS Diagnostics d.o.o. as a software developer. MM and DG declare no potential conflict of interest.

### Additional information

**Correspondence** and requests for materials should be addressed to V.G.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025