Wechsler Test of Adult Reading in Parkinson's: a stable yet imperfect measure of premorbid cognitive function

Kyla-Louise Horne, PhD[1], Reza Shoorangiz, PhD[1], Daniel J. Myall, PhD[1], Toni L. Pitcher, PhD[1,2], Tim J. Anderson, FRACP, MD[1,2,3], John C. Dalrymple-Alford, PhD[1,2,4], & Michael R. MacAskill, PhD[1,2]

[1] New Zealand Brain Research Institute, 66 Stewart St, Christchurch, New Zealand
[2] Department of Medicine, University of Otago, Christchurch; Christchurch, New Zealand
[3] Department of Neurology, Christchurch Hospital, Christchurch, New Zealand
[4] School of Psychology, Speech, and Hearing, University of Canterbury, Christchurch, New Zealand

Author Note

Correspondence concerning this article should be addressed to Kyla-Louise Horne, PhD, 66 Stewart St, Christchurch 8011, New Zealand. E-mail: kyla.horne@nzbri.org

Abstract

Purpose. To assess long-term trajectories of Wechsler Test of Adult Reading (WTAR) scores in people with Parkinson's as a function of age. The WTAR has been recommended as a measure of premorbid cognitive function in English speakers with Parkinson's disease and as a reference against which to assess current cognitive status. For this, however, it needs to be shown that WTAR scores remain stable despite the substantial cognitive deterioration that can occur in Parkinson's. Methods. From 252 Parkinson's and 57 Control participants who had completed at least two WTARs, we analyzed scores over time using latent class trajectory modeling. This allows for individual participants to be classified into data-driven clusters, depending on the shape of their longitudinal trajectory. Results. WTAR scores were quite stable within both Controls and Parkinson's participants, even for those who progressed to dementia. This validates it as a research tool for comparing premorbid function at a group level. In both Parkinson's and Controls, and regardless of current cognitive status, the distribution of scores was, however, higher than expected from the population norms, making it an unreliable benchmark against which to detect cognitive decline at an individual level. Conclusion. The WTAR is stable in Parkinson's even when participants decline from normal cognitive function to early dementia. Nonetheless, its apparent over-estimation of premorbid IQ and the impracticality of implementing analogous tests in many other languages makes it poorly-suited for detecting current cognitive impairment in individuals with Parkinson's.

Keywords: Parkinson disease, cognitive impairment, dementia, neuropsychology, premorbid function

Word count: Abstract: 239 (limit 250), main body: 3870 (limit 4000)

Wechsler Test of Adult Reading in Parkinson's: a stable yet imperfect measure of premorbid cognitive function

Introduction

The prevalence of Parkinson's is rising,[1,2] and hence the burden of dementia associated with the disease will also continue to grow. There is therefore value in detecting the transitional stage of mild cognitive impairment in Parkinson's, in particular because it may provide a therapeutic window for future treatments, prior to the irreversible pathological damage associated with dementia. To provide for consistent delineation of this transitional state, the International Parkinson and Movement Disorder Society (MDS) published guidelines in 2012 for the diagnosis of mild cognitive impairment in Parkinson's (PD-MCI).[3]

When the diagnosis is based on a comprehensive battery of multiple individual neuropsychological tests (rather than on a single scale of global cognitive ability), the MDS guidelines propose that significant impairment may be determined in three ways. Firstly, current performance on a test can be shown to be below appropriate norms (which may be established either relative to standardized norms for that test, or relative to a local matched control group). This approach has the disadvantage that the cause of poor current performance cannot be distinguished between that due to a recent decline (such as the onset of MCI or dementia), or a long-standing poor level of cognitive ability. To unambiguously detect recent impairment in cognitive status therefore requires some way of demonstrating a *decline* in function within an individual, rather than simply poor current performance. To achieve this, the guidelines suggest seeking evidence of either a decline in performance on serial testing, or, in the absence of such prior testing, evidence of a decline from an individual's *estimated* premorbid level of function. For pragmatic reasons, however, methods based upon current norms still tend to predominate in research studies.[4] In practice, one seldom has reference to prior formal serial testing. Even in a well-resourced

research setting, serial testing remains somewhat ambiguous, as the first available test session may itself be contaminated by early disease-related cognitive impairment. This would produce a falsely low baseline and thereby impede the ability to detect cognitive decline.

A reliable method of estimating premorbid cognitive function is therefore appealing. Establishing a person's premorbid ability addresses the shortcomings of both the current-norms approach (by showing whether current poor performance is indeed a decline from past status), and the issue with the timing of serial testing (which will usually begin only after disease diagnosis). The Wechsler Test of Adult Reading (WTAR)[5] was suggested in the MDS guidelines as being suitable for this purpose (along with the National Adult Reading Test, NART).[6] Both tests are based on the ability to correctly pronounce phonetically-irregular English words, such as 'porpoise' or 'hyperbole'. Unlike many neuropsychological functions, this ability has been shown to be preserved in the face of both normal ageing and a broad range of brain insults.[5,6]

When the WTAR was initially normed, it was tested in Parkinson's as well as a number of other neurological or psychiatric disorders, such as Huntington's, schizophrenia, and traumatic brain injury.[5] Only in Alzheimer's disease were scores shown to be significantly lower than in matched controls.[5,6] Additionally in the Alzheimer's group, WTAR scores were lower in those with lower overall current cognitive status,[6] further indicating that it is not a stable measure of premorbid function in that condition. Those results were, however, published in 2001, well before the formal MDS guidelines on diagnosing PD-MCI[3] and Parkinson's dementia[7] were promulgated. It is therefore unknown to what extent the findings from that sample are valid in the face of the significant cognitive deterioration that can occur in Parkinson's, as it could not have been formally diagnosed at the time. Moreover, their Parkinson's sub-sample was also comprised of only 10 people: given the marked heterogeneity of function in this disorder, this restricts the validity of generalizing from those findings. We know of only one study in which the

WTAR has actually been used as a criterion against which to establish mild cognitive impairment in Parkinson's.[8] The proportion classified as PD-MCI using WTAR-estimated premorbid functioning was strikingly higher than when using the more conventional method of comparing current performance against norms (79% vs 33%). This naturally leads to questions about whether the WTAR is in fact a valid measure to be used for cognitive classification in Parkinson's. In particular, if using the WTAR leads to most people with Parkinson's being classified as having MCI, does that result in a category that no longer has any prognostic or discriminative utility?

Performance on the WTAR is known to be impacted negatively by Alzheimer's dementia, but as noted the initial claim of stability in Parkinson's was based on very limited evidence.[5] We therefore sought to examine WTAR scores longitudinally in a large and cognitively formally-characterized cohort that spanned the range from normal function through to mild cognitive impairment and dementia. We included only those participants who had completed a minimum of two WTARs, so that the trajectory of the measure within individuals over time could be examined. To do this, we used latent class trajectory analysis[9]. This allows for detecting heterogeneous patterns of change over time in a data-driven fashion, rather than assigning individuals to *a priori* sub-groups. That is, we combined both Control and Parkinson's participants in a single pool. The modeling process examines individual subjects and assigns them into clusters depending on similarities in their longitudinal trajectories. If the presence of either Parkinson's or cognitive impairment affects longitudinal WTAR scores, this should be reflected by differential membership across groups in the resulting data-driven trajectory clusters. For example, Parkinson's participants with normal cognition might cluster together with most of the Controls in a "flat" trajectory class, while those showing cognitive impairment might cluster in one or more trajectory classes showing decline over time. Conversely, if the WTAR is truly stable in people with Parkinson's, we would expect them to cluster together with controls in relatively flat trajectory classes, regardless of their current level of

cognitive function.In particular, we sought to examine whether WTAR performance declines in the dementia due to Parkinson's, as it does in Alzheimer's. If so, this would impact its usefulness in estimating premorbid function in the way proposed in the MDS guidelines for cognitive diagnosis.

Methods

Participants

The New Zealand Parkinson's Progression Programme (NZP[3]) is a longitudinal study of a convenience prevalence sample of idiopathic Parkinson's participants,[10] largely recruited from the specialist Movement Disorders Clinic at the New Zealand Brain Research Institute (NZBRI). Ethical approval was granted by the Southern Health and Disability Ethics Committee of the New Zealand Ministry of Health. The ongoing recruitment commenced in 2007, and includes patients ranging from the recently-diagnosed to those with advanced disease. Inclusion and exclusion criteria have been described previously.[11] For this analysis, we selected data from the 252 Parkinson's and 57 Control participants who had completed at least two WTARs (range 2 – 8, mean = 3.70, total number of measures = 1155). The period between successive WTAR assessments was multi-modal, mostly clustering around intervals of one and two years, due to varying follow-up periods over the course of the wider study (see Supplementary Material). Of the Parkinson's participants, the mean duration since diagnosis at the first WTAR assessment was 5.0 years (range 0 – 23) and the mean duration of follow-up was a further 4.6 years (range 0 – 11). All participants were classified as having normal cognition, mild cognitive impairment, or dementia via a comprehensive Level II neuropsychological battery,[11,12] administered in accordance with MDS guidelines.[3,7] A PD-MCI classification was based on at least two test scores falling ≥ 1.5 standard deviations below norms in at least one cognitive domain, without significant impairment in activities of daily living. For PD dementia, a significant

impairment ($\geq$ 2.0 standard deviations below normative data) in at least two cognitive domains was required, as well as evidence of significant impairment in activities of daily living.[11,12] The WTAR itself was not used in the cognitive diagnostic classification procedures. Characteristics of the sample are shown in Table 1. All participants were speakers of New Zealand English, and were assessed against the US norms of the WTAR, corrected for age, sex, and years of education. The provided ethnicity classifications are not applicable in a New Zealand context and all participants were assessed against norms for "Whites". Of the 249 Parkinson's participants with a recorded ethnicity, 240 (96.4%) identified as Pākehā (New Zealand European), 2 as Māori, 1 as Samoan, 1 as Indian, and 5 were 'Other'. Of the Controls, 56 (98.2%) identified as Pākehā and 1 was 'Other'.

Latent class trajectory modeling

We fitted latent class trajectory models using the *hlme* function from the *lcmm* package[9] (version 2.0.0), running in the R statistical environment[13] (version 4.2.1), and used the *tidyverse* constellation of packages[14] for data manipulation and vizualisation. The dependent variable was WTAR-estimated premorbid IQ, initially modeled within each subject as a polynomial (cubic) function of age, to allow for non-linear trajectories. The resulting trajectories were close to straight lines (see Supplementary Material) so we simplified the models to be a linear function of age. The models were not informed by any other variables (such as the diagnostic categories of PD-MCI or dementia). We used age rather than disease duration as the time metric to allow for direct comparison against any aging effect in controls.

A latent class model is fitted by first specifying an *a priori* number of classes (i.e. clusters) into which the participants can be divided. For example, if just one cluster is specified, then the trajectory that is produced will simply describe the average change in score over time for all participants, disregarding any possible sub-groups. If two classes are specified, then the model will determine two trajectories that optimally separate the total

sample into two clusters, on the basis of differing performance over time. Multiple independent models are fitted, with increasing numbers of classes. Those models are then compared to see which produces the best description of the data. The classes are termed 'latent' as they are not known *a priori*, but instead arise from the data. The resulting classes can, however, then be compared to known groupings (such as Parkinson's vs Controls). This allows us to examine to what extent performance over time might be driven by known factors. For example, if Control and Parkinson's participants perform differently over time, they should fall unequally into the various latent classes (for models with more than one class).

We fitted five separate models, with the specified number of latent classes ranging from 1 to 5. The *gridsearch* function of the *lcmm* package was used to run 500 departures from random initial values for each multi-class model, using the one-class model from which to generate the starting values. The maximum number of iterations within the *hlme* function was set at 1000. Parameters corresponding to the best log-likelihood were used as initial values for the final estimation of the parameters[9]. The optimal model of the five candidates was selected on the basis of it converging and having the lowest BIC (Bayesian information criterion). That is, this allowed us to determine whether longitudinal performance on the WTAR was best described by the participants falling into either 1, 2, 3, 4, or 5 underlying trajectory classes.

When reporting values from the chosen model, we formed the interval in brackets following the maximum likelihood estimate (MLE) by subtracting and adding 1.96 times the standard error given by *hlme* to the MLE.

Reproducibility

The code and anonymised dataset extracts sufficient to reproduce the analyses and generate this manuscript are publicly available at github.com/nzbri/wtar-trajectory. There

are a number of decisions that can affect the outcome of a latent class modeling analysis and for reproducibility in the Supplementary Material we therefore report our performance against the 16-item checklist "Guidelines for Reporting on Latent Trajectory Studies".[15]

Results

Distribution of scores

The distributions of all WTAR-estimated IQ scores for the Parkinson's and Control participants are shown in Figure 1. The mean of the latest score for participants in each of the groups is given in Table 1, with both being well above the expected population norm mean IQ score of 100. The population distribution of IQ should be symmetrical about 100, yet in our sample, 87.0% of Parkinson's WTAR scores were greater than or equal to 100, as were 92.5% of Control scores.

Longitudinal trajectories

The raw data showed that WTAR scores within individuals were relatively constant over time (Figure 2A). This was evident even in participants whose overall cognitive performance declined substantially, as they progressed to PD-MCI or dementia (Figure 2B). We modeled WTAR scores as a linear function of age, with repeated measures within individuals, and fitted candidate models that ranged from having one to five latent trajectory classes. We selected the two-class model on the basis of it having the lowest BIC value. Full details on the modeling process are provided in the Supplementary Material.

Individuals were assigned to the class for which they had the highest posterior classification probability. The mean assigned probability was 0.91 (SD 0.14), with a value of 0.7 being regarded as acceptable model performance.[16] The two trajectories are depicted superimposed upon the raw data in Figure 2A. We labeled the first trajectory 'Typical', as this class captured 78% of all participants (81% of the PD and 65% of Control

participants). The intercept of the Typical trajectory, representing the estimated WTAR score of a person at the mean age of assessment (70.7 years) was 107.6 [105.9, 109.2]. Scores then gently declined at a rate of -1.2 [-1.9, -0.6] IQ points per decade. The second trajectory we labeled 'High performers'. This class, capturing the remaining 22% of all participants, had an intercept of 121.0 [118.9, 123.1] and a slope of -0.2 [-1.2, 0.8] points per decade (effectively flat). That is, both trajectories were relatively stable over time, and were separated primarily due to the initial scores of their respective clusters of participants. We make no strong claims as to the number of classes. For example a single trajectory model, somewhat intermediate between the two reported here, would also describe the data relatively well, and a single trajectory was the optimal model when cubic rather than linear trajectories were specified (see Supplementary Material). The key findings are that the trajectories are relatively flat, and do not separate performance between the Parkinson's and Control participants, nor between people with preserved versus impaired cognitive function.

Discussion

Unlike the progression to Alzheimer's dementia, we found that WTAR scores were quite stable over time in Parkinson's, despite many of our participants undergoing substantial decline in overall cognitive performance over the course of the study, including to dementia (Figure 2B). Latent class analysis revealed that in our sample people were best classified into two trajectory classes. These appeared to be largely driven by two clusters of performance at baseline, rather than being trajectories that substantially differed longitudinally. Those with very high initial WTAR scores (the 'High performers' class) maintained high scores (Figure 2A), even if they subsequently developed dementia (Figure 2B). The remaining 78% of participants (the 'Typical' class) had a wider range of initial scores, but were again relatively stable, showing an estimated decline of only -1.2 points per decade. If the WTAR was used to assess disparities between current cognitive

performance and a premorbid level of function, a decadal decline of that magnitude would be unlikely to influence assignment to clinical categories like MCI and dementia.

Measures of premorbid function are closely tied to the concept of cognitive reserve.[19] For example, higher NART scores have been associated with better current neuropsychological test performance in people with Parkinson's,[20] indicating that higher cognitive reserve may be protective against subsequent cognitive decline. This could, however, lead to circular reasoning if the premorbid measure was not reliable, because if the disease process also reduces current performance on that measure, then inferring poorer cognitive reserve would not be valid. That is, due to current impairments, we would be estimating a lower level of premorbid function than the person actually had. Our findings, of stable WTAR scores despite substantial current cognitive decline, provide support for the WTAR being a stable indicator of cognitive reserve in Parkinson's.

We retain reservations about the use of the WTAR, however. Although still in wide use, it is outdated, inasmuch as it was originally devised to estimate premorbid IQ scores from the WAIS-III (Wechsler Adult Intelligence Scale III, released in 1997).[5] The WAIS-IV was released in 2008) and is now in widespread use, while the WAIS-V is currently under development. The TOPF (Test of Premorbid Function) estimates WAIS-IV IQ and was designed as a successor to the WTAR, but has had limited uptake in the Parkinson's literature.[21] This perhaps reflects the inertia of substantial research studies (including our own), in which investigators have a natural desire to retain backwards compatibility with measures that they have gathered over a long period of time. The age of an IQ-related test is particularly important. The Flynn effect[22] is the robust finding that cohort IQ scores have increased substantially over time (by up to 1 standard deviation per generation). Our sample would (on average) have been born well after their age-matched WAIS-III normative sample group and that might partly explain their above-average estimated IQ levels. Another possibility for the high scores that we observed is that those who volunteer for research are unlikely to be representative of their respective populations. Such a

self-selection effect, however, applies more to the controls than to the Parkinson's group, particularly because our controls were excluded (at study entry only) if they showed cognitive impairment. This restriction did not apply to the Parkinson's participants, but again, as research volunteers, they may not be representative of the Parkinson's population at large. As shown in Table 1, the mean education level for both the Controls and Parkinson's participants was close to that of a complete New Zealand high school education (13 years). Although a selection bias is undoubtedly a factor in the NZP[3] sample, we do not believe it could account for the predominance of estimated IQ scores above the putative population norm of 100 (87.0% of Parkinson's scores and 92.5% of Control scores).

To provide evidence of cognitive decline, scores on the tests from a neuropsychological test battery must fall below some reference value. If the WTAR was used as an individualized premorbid reference for those scores, rather than a per-test z-score norm of zero, we would expect to see inflated numbers of participants being classified as MCI, because of the over-estimation of premorbid function. The PD-MCI guidelines do not actually formalize how the results of premorbid function tests should be used in practice to demonstrate decline in cognitive functioning.[23,24] We are aware of only one study[8] that has compared the norms-based vs premorbid estimation approaches to establishing cognitive impairment. They converted the WTAR-estimated full-scale IQ to a z-score and used that as the reference against which to test each individual neuropsychological test z-score. Any test of current performance that was lower by more than 1.5 SD from the estimated premorbid level was deemed impaired. Of their sample of 139 non-demented people with Parkinson's, as noted earlier, the WTAR method classified many more of them as having PD-MCI (79%) than did a norms-based classification method (33%). Marras *et al.* interpreted this as evidence that most people with Parkinson's have undergone at least some decline relative to their premorbid function, with the WTAR providing a sensitive method to detect it. Conversely, however, that finding might support our contention that the WTAR simply over-estimates a person's premorbid

level of function – just as in our sample, where the distribution of estimated IQ values was skewed well above the supposed population mean of 100.

The WTAR (and similar tests such as the NART and TOPF) are also limited in that they are inherently tied to peculiarities of written and spoken English. The orthography (writing system) of a language can be described as deep/opaque when it contains many irregularities in its mappings of graphemes to phonemes, or shallow/transparent when the mappings are largely regular.[25] In a quantitative analysis of 17 orthographies,[26] the reading of English words stood out as particularly opaque, whereas by comparison Arabic, Finnish, Korean, Serbo-Croatian, and Turkish were highly transparent. Attempting to devise analogues of tests like the WTAR for speakers of languages with transparent orthographies is unlikely to be fruitful.

Even across varieties of English, both between and within countries, the US- and UK-based selection of words can be problematic. For example, the TOPF was found to not reliably predict WAIS-IV IQ in New Zealand English speakers of Māori ethnicity.[27] A word like 'porpoise', ranked in the TOPF as relatively simple (28th of 70 items), resulted in more errors than a word like 'plethora' (ranked 42nd of 70), likely reflecting different word frequencies between North American and New Zealand English. After the commencement of our study, a New Zealand Adult Reading Test (NZART) was created[28], with a more culturally-relevant selection of words, but this will require further development to supplant the established international tests. Neuropsychological screening tests developed in the "Anglosphere" have been shown to have poor cross-cultural validity in Parkinson's.[29] This extends even to visuoperceptual and non-verbal executive tasks previously assumed to be "culture fair".[30] It is perhaps therefore not surprising that an inherently language-based test faces issues with cross-cultural applicability. In subsequent revisions of the PD-MCI diagnostic criteria, if premorbid measures continue to be advocated, more consideration could be given to alternative techniques[31] that could have improved validity in a wider international and cross-cultural context.

In summary, the WTAR is a measure that is stable over time in people with Parkinson's, even in the face of a marked decline in current cognitive performance. This makes it a valid research tool to probe cognitive reserve and premorbid function differences across groups. In absolute terms, however, it likely overestimates premorbid IQ, making it unsuitable to use as a benchmark to establish a decline in function within individuals. All such measures likely suffer from cross-cultural word choice issues, even across different varieties of English. More importantly, analogous tests are not even able to be created for languages that do not share the orthographic peculiarities of English. The strikingly stable recall of atypical pronunciation should not be thought of as a fundamental neuropsychological function, but rather a phenomenon that is only evident in particular linguistic and cultural populations. Therefore, in future revisions of the international guidelines for classifying cognitive impairment in Parkinson's, we believe that a test of premorbid function like the WTAR should no longer be recommended as one of the means of formally establishing cognitive decline at an individual level. Due to its remarkable stability in Parkinson's despite substantial cognitive impairment, this and related measures will, however, remain a valuable research tool in the Anglosphere, at least for testing groups for differences on premorbid functioning. The absolute estimation of premorbid function in Parkinson's should be determined for newer tests such as the TOPF and its successors. With newer norms, they may address the over-estimation we found with the WTAR.

References

1 Myall DJ, Pitcher TL, Pearson JF, *et al.* Parkinson's in the oldest old: Impact on estimates of future disease burden. *Parkinsonism & Related Disorders* 2017;**42**:78–84. doi:10.1016/j.parkreldis.2017.06.018

2 Pitcher TL, Myall DJ, Pearson JF, *et al.* Parkinson's disease across ethnicities: A nationwide study in New Zealand. *Movement Disorders* 2018;**33**:1440–8. doi:10.1002/mds.27389

348 3 Litvan I, Goldman JG, Tröster AI, *et al.* Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines.
349 *Mov Disord* 2012;**27**:349–56. doi:10.1002/mds.24893

350 4 Aarsland D, Batzu L, Halliday GM, *et al.* Parkinson disease-associated cognitive
351 impairment. *Nat Rev Dis Primers* 2021;**7**:47. doi:10.1038/s41572-021-00280-3

352 5 Psychological Corporation. *WTAR Wechsler Test of Adult Reading Manual.* San
353 Antonio: : Harcourt Assessment 2001.

354 6 Strauss E, Sherman EMS, Spreen O. *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary.* 3rd ed. Oxford: : Oxford University
355 Press 2006.

356 7 Dubois B, Burn D, Goetz C, *et al.* Diagnostic procedures for Parkinson's disease dementia: Recommendations from the Movement Disorder Society Task Force. *Mov*
357 *Disord* 2007;**22**:2314–24.

358 8 Marras C, Armstrong MJ, Meaney CA, *et al.* Measuring mild cognitive impairment in patients with Parkinson's disease: Measuring PD-MCI. *Mov Disord* 2013;**28**:626–
359 33. doi:10.1002/mds.25426

360 9 Proust-Lima C, Philipps V, Liquet B. Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *J Stat Softw* 2017;**78**:1–56.
361 doi:10.18637/jss.v078.i02

362 10 MacAskill MR, Pitcher TL, Melzer TR, *et al.* The New Zealand Parkinson's Progression Programme. *Journal of the Royal Society of New Zealand* 2022;1–23.
363 doi:10.1080/03036758.2022.2111448

364 11 Wood K-L, Myall DJ, Livingston L, *et al.* Different PD-MCI criteria and risk of dementia in Parkinson's disease: 4-year longitudinal study. *npj Parkinson's Dis*
365 2016;**2**:15027. doi:10.1038/npjparkd.2015.27

12 Dalrymple-Alford JC, Livingston L, MacAskill MR, *et al.* Characterizing mild cognitive impairment in Parkinson's disease. *Mov Disord* 2011;**26**:629–36. doi:10.1002/mds.23592

13 R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: : R Foundation for Statistical Computing 2021. https://www.R-project.org/

14 Wickham H, Averick M, Bryan J, *et al.* Welcome to the Tidyverse. *J Open Source Softw* 2019;**4**:1686. doi:10.21105/joss.01686

15 Schoot R van de, Sijbrandij M, Winter SD, *et al.* The GRoLTS-checklist: Guidelines for reporting on latent trajectory studies. *Struct Equ Modeling* 2017;**24**:451–67. doi:10.1080/10705511.2016.1247646

16 Lennon H, Kelly S, Sperrin M, *et al.* Framework to construct and interpret latent class trajectory modelling. *BMJ Open* 2018;**8**:e020683. doi:10.1136/bmjopen-2017-020683

17 MacAskill MR. Distribution of Wechsler Test of Adult Reading (WTAR) estimated premorbid IQ scores in Parkinson's disease and controls. 2022. doi:10.6084/M9.FIGSHARE.19126538.V1

18 MacAskill MR. Longitudinal WTAR-estimated premorbid IQ score trajectories in Parkinson's disease and controls. 2022. doi:10.6084/M9.FIGSHARE.19126553

19 Stern Y. Cognitive reserve in ageing and Alzheimer's disease. *Lancet Neurol* 2012;**11**:1006–12. doi:10.1016/S1474-4422(12)70191-6

20 Koerts J, Tucha L, Lange KW, *et al.* The influence of cognitive reserve on cognition in Parkinson's disease. *J Neural Transm* 2013;**120**:593–6. doi:10.1007/s00702-012-0916-6

21 Bright P, Linde I van der. Comparison of methods for estimating premorbid intelligence. *Neuropsychol Rehabil* 2020;**30**:1–4. doi:10.1080/09602011.2018.1445650

386 22 Flynn JR. Searching for justice. The discovery of IQ gains over time. *Am Psychol* 1999;**54**:5–20. doi:10.1037/0003-066X.54.1.5

388 23 Marras C, Tröster AI, Kulisevsky J, *et al.* The tools of the trade: A state of the art 'How to Assess Cognition' in the patient with Parkinson's disease. *Mov Disord* 2014;**29**:584–96. doi:10.1002/mds.25874

390 24 Geurtsen GJ, Hoogland J, Goldman JG, *et al.* Parkinson's disease mild cognitive impairment: Application and validation of the criteria. *J Parkinson Dis* 2014;**4**:131–7. doi:10.3233/JPD-130304

392 25 Frost R, Katz L, Bentin S. Strategies for visual word recognition and orthographical depth: A multilingual comparison. *J Exp Psychol Hum Percept Perform* 1987;**13**:104–15. doi:10.1037/0096-1523.13.1.104

394 26 Marjou X. OTEANN: Estimating the transparency of orthographies with an artificial neural network. In: *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*. Online: : Association for Computational Linguistics 2021. 1–9. doi:10.18653/v1/2021.sigtyp-1.1

396 27 Dudley M, Scott K, Barker-Collo S. Is the test of premorbid functioning a valid measure for Māori in New Zealand? *N Z J Psychol* 2017;**46**:8.

398 28 Starkey NJ, Halliday T. Development of the New Zealand Adult Reading Test (NZART): Preliminary findings. *N Z J Psychol* 2011;**40**:13.

400 29 Statucka M, Cherian K, Fasano A, *et al.* Multiculturalism: A challenge for cognitive screeners in Parkinson's disease. *Mov Disord Clin Prac* 2021;**8**:733–42. doi:10.1002/mdc3.13240

402 30 Statucka M, Cohn M. Origins matter: Culture impacts cognitive testing in Parkinson's disease. *Front Hum Neurosci* 2019;**13**:269. doi:10.3389/fnhum.2019.00269

404 31 Franzen MD, Burgess EJ, Smith-Seemiller L. Methods of estimating premorbid functioning. *Arch Clin Neuropsychol* 1997;**12**:711–38. doi:10.1093/arclin/12.8.711

405

Table 1

Demographics of the Parkinson's and Control groups, and cognitive status as at their latest assessment.

	Control	Parkinson's
n	57	252
Male	63.2%	69.4%
Age (SD)	76.1 (7.9)	72.7 (7.0)
Years of education (SD)	13.6 (2.7)	12.7 (2.6)
WTAR (SD)	112.6 (9.5)	108.9 (9.3)
Normal cognition	94.7%	41.3%
MCI	3.5%	34.9%
Dementia	1.8%	23.8%

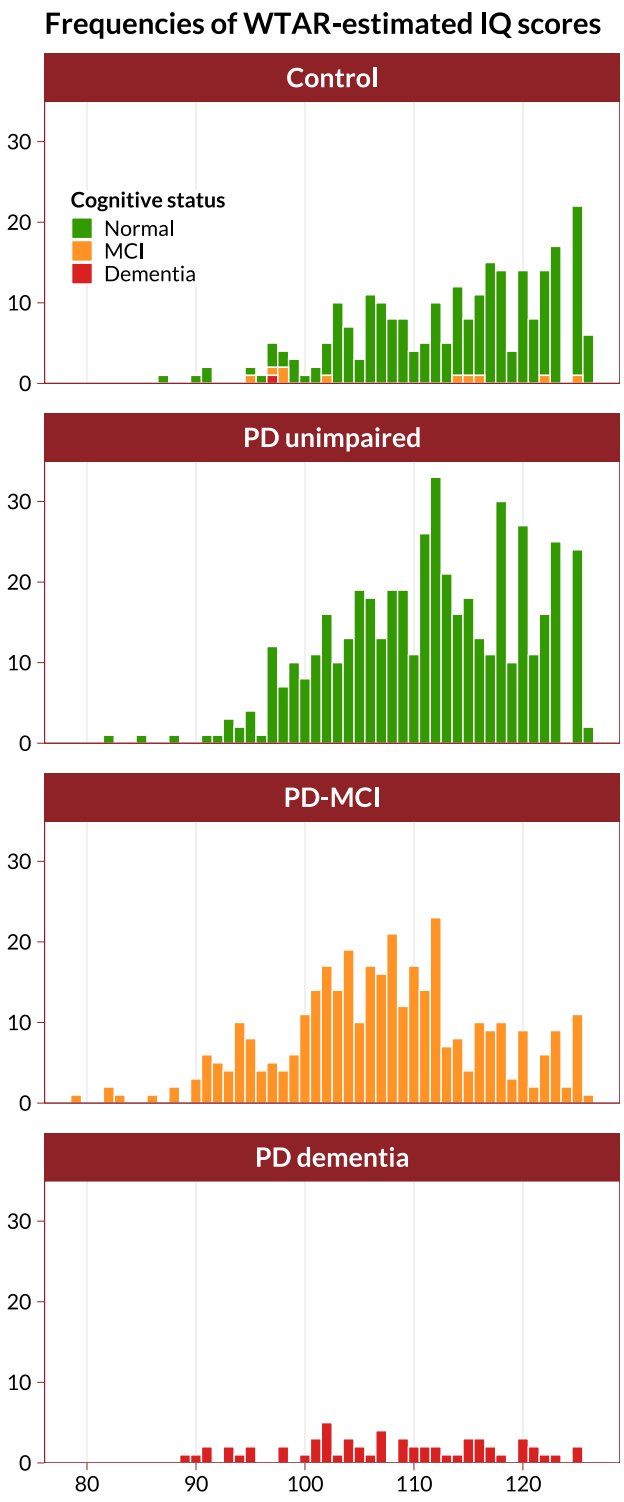

Figure 1. The distribution of WTAR-estimated IQ scores across all participants (containing at least two observations from each). Observations are color-coded by the participant's cognitive status at each session. Regardless of current cognitive status, WTAR scores span a wide range. For each cognitive classification, the center of distribution is shifted well above the population norm mean of 100. Figure available under an open CC-BY licence.[17]

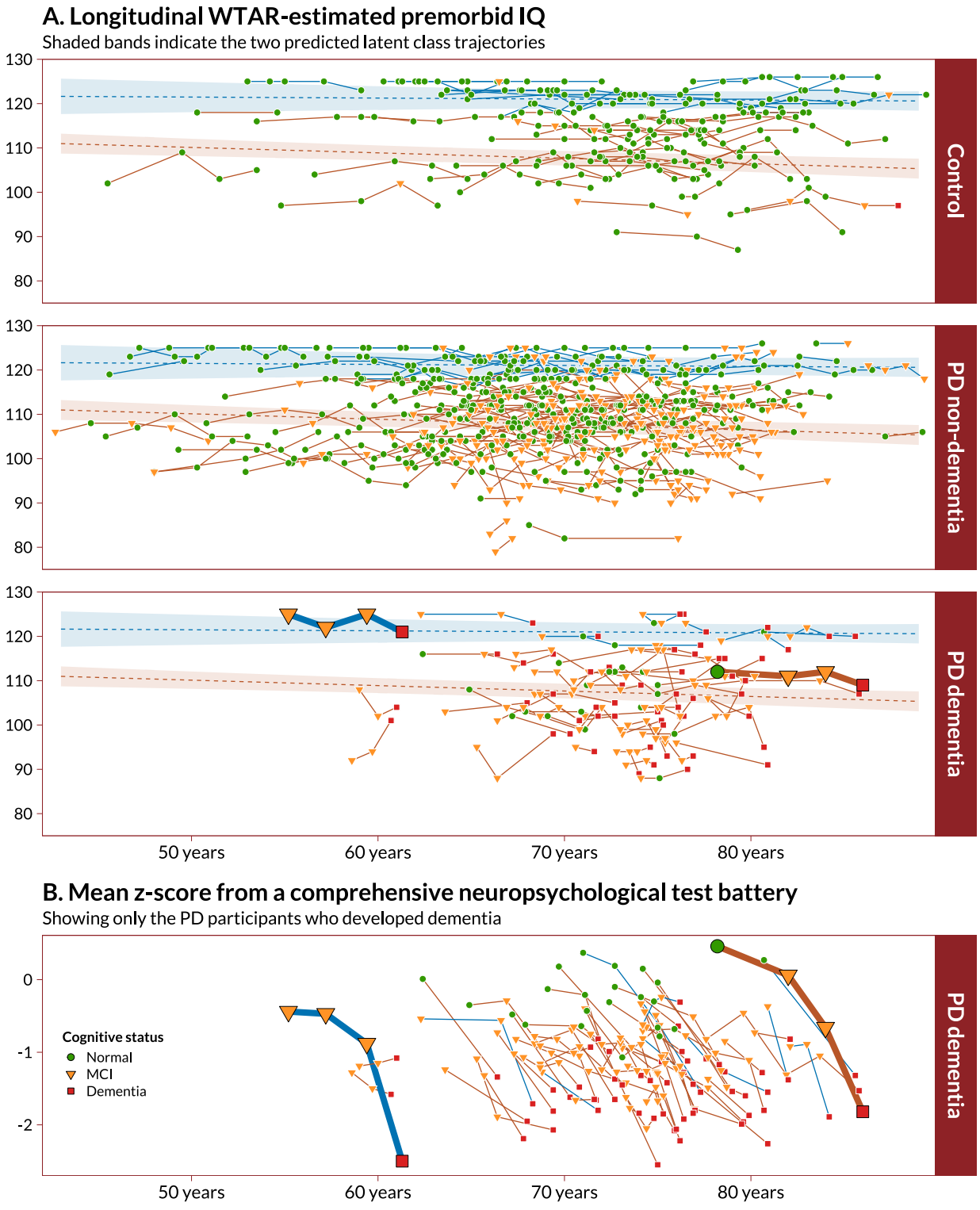

Figure 2. ***(A)*** Longitudinal WTAR score trajectories for all individuals, divided into Controls, Parkinson's participants who remained dementia-free, and those who developed dementia. Superimposed are the two predicted latent class trajectories (blue ribbon = *High performers*, brown ribbon = *Typical*). Blue and brown connecting lines link observations from individuals from each class. ***(B)*** Mean z-score from multiple tests in a comprehensive

Appendix

Supplementary Material

Horne *et. al.*, *"Wechsler Test of Adult Reading in Parkinson's: a stable yet imperfect measure of premorbid cognitive function."*

Here we report the analysis of the above paper against the following checklist for reporting latent class trajectory analyses:

van de Schoot, R., Sijbrandij, M., Winter, S. D., Depaoli, S., & Vermunt, J. K. (2017). The GRoLTS-checklist: Guidelines for reporting on latent trajectory studies. *Structural Equation Modeling: A Multidisciplinary Journal, **24,*** 451–467. https://doi.org/10.1080/10705511.2016.1247646

1. Is the metric of time used in the statistical model reported?

We used age as the time metric (in years to one decimal place). This was to allow for direct comparison against controls (which would not be possible with disease duration). Follow-up time within the study was also not considered, as participants can enter the study at arbitrary points (in terms of both age and disease duration).

To avoid instabilities in the cubic models, age was standardised by subtracting the global mean age at assessment and dividing by ten. For comparability, this was also done for the linear models. In the visualisations, predictions were projected back into untransformed age in years.

2. Is information presented about the mean and variance of time within a wave?

The metric was age rather than intended study time points. The overall goal in the study was to follow-up participants two-yearly, however, the interval between follow-up times was multi-modal and varied both within and across participants (see Figure A1).

This variability was due to both changing study demands (such as recruiting participants for sub-studies) and participant health and availability.

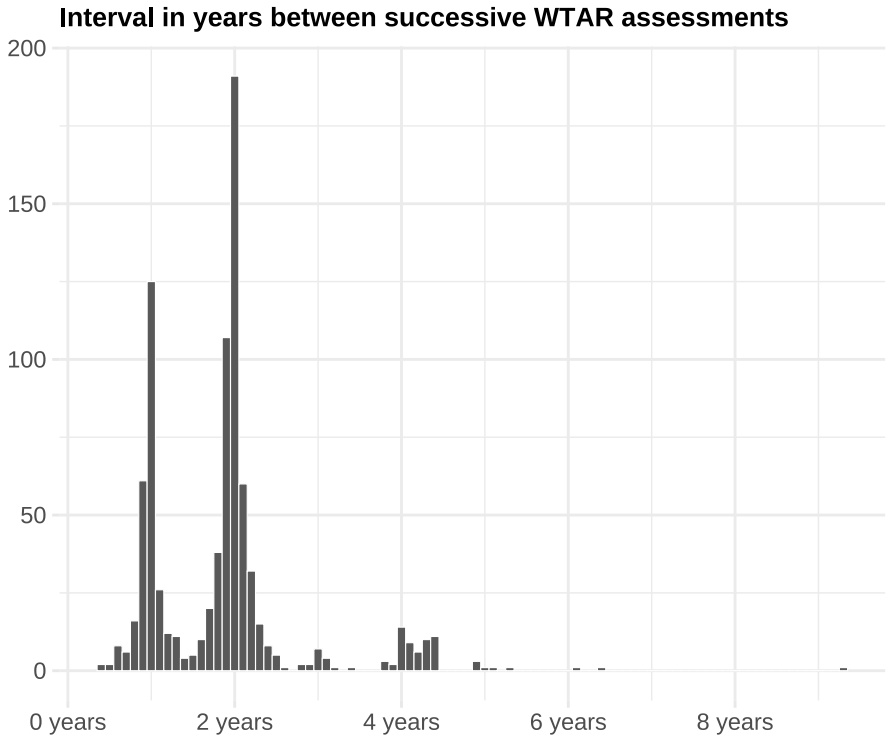

Figure A1. Frequencies of the interval between successive WTAR assessments.

3a. Is the missing data mechanism reported?

Data was not missing at random: participants are more likely to drop out of the study due to increasing disease severity, other health issues, and old age. As shown in Figure 2, however, many remained in the study until the end point of WTAR collection (the onset of dementia). Those participants did not show a decline in WTAR performance relative to other participants. Due to the findings of this study (that the WTAR is relatively stable throughout the disease course), differential drop-out is unlikely to affect the results.

3b. Is a description provided of what variables are related to attrition/missing data?

They are speculated upon in *3a* above.

3c. Is a description provided of how missing data in the analyses were dealt with?

As noted in *3a* above, missing data was not thought to be problematic in this analysis and was not included explicitly in the modeling process.

4. Is information about the distribution of the observed variables included?

In Figure **??** in the accompanying manuscript, we show the distribution of the WTAR-estimated IQ scores in each group, showing the cognitive status at each observation.

5. Is the software mentioned?

See the Methods section in the accompanying manuscript.

6a. Are alternative specifications of within-class heterogeneity considered (e.g., LGCA vs. LGMM) and clearly documented? If not, was sufficient justification provided as to eliminate certain specifications from consideration?

An LGMM which includes a varying intercept and slope by participant was used in this analysis and, given the nature of the within-subject variance in the data, is more appropriate than an LGCA.

6b. Are alternative specifications of the between-class differences in variance–covariance matrix structure considered and clearly documented? If not, was sufficient justification provided as to eliminate certain specifications from consideration?

The variance-covariance matrix was allowed to vary by a class-specific proportional parameter. The more restrictive case of restricting all classes to a common variance-covariance matrix structure was not considered.

7. Are alternative shape/functional forms of the trajectories described?

We initially modeled the trajectories using a cubic function, to allow for the depiction of any curvilinear decline in performance. The resulting trajectory in the one-class cubic model (which had the lowest BIC) was effectively flat, and hence we moved to linear modeling for the ease of interpretation of the model coefficients.

8. If covariates have been used, can analyses still be replicated?

Not applicable: age was the only predictor in the models.

9. Is information reported about the number of random start values and final iterations included?

The `gridsearch()` function of the `lcmm` package was used to run 500 departures from random initial values for each multi-class model, using the one-class model to generate those starting values from. The maximum number of iterations within the `hlme()` function was set at 1000.

10. Are the model comparison (and selection) tools described from a statistical perspective?

We selected the two-class linear model on the basis of it converging and having the lowest BIC value.

11. Are the total number of fitted models reported, including a one-class solution?

For both linear and cubic trajectories, models were fitted with one through five classes. See the model comparison tables in Tables A1 and A2.

[tbp]

Table A1

Comparison of hlme models with G = 1 to 5 linear latent trajectory classes.

G	loglik	converged	BIC	entropy	%class1	%class2	%class3	%class4	%class5
1	-3,116.9	yes	6,268.2	1.00	100.0				
2	-3,102.4	yes	6,262.2	0.71	78.3	21.7			
3	-3,092.6	yes	6,265.5	0.64	44.0	40.5	15.5		
4	-3,086.2	yes	6,275.5	0.69	30.4	9.7	46.9	12.9	
5	-3,082.3	yes	6,290.8	0.68	27.5	48.9	15.5	2.3	5.8

[tbp]

Table A2

Comparison of hlme models with G = 1 to 5 cubic latent trajectory classes.

G	loglik	converged	BIC	entropy	%class1	%class2	%class3	%class4	%class5
1	-3,108.2	yes	6,302.4	1.00	100.0				
2	-3,093.6	yes	6,307.6	0.77	81.2	18.8			
3	-3,083.0	yes	6,320.9	0.75	61.8	13.3	24.9		
4	-3,076.4	yes	6,342.0	0.68	23.6	23.9	22.7	29.8	
5	-3,069.0	yes	6,361.7	0.71	21.7	3.2	18.8	24.6	31.7

[tbp]

Table A3

Number within each class in the selected model.

	Typical	High
N	242	67
%	78.32	21.68

12. Are the number of cases per class reported for each model (absolute sample size, or proportion)?

The number of cases within each class of the selected model is shown in Table A3. See Tables A1 and A2 for the percentage of cases within each class in all candidate models.

13. If classification of cases in a trajectory is the goal, is entropy reported?

Classification of cases was not a goal of the analysis. However, entropy-related measures are shown in Tables A4 and A5.

[tbp]

Table A4

Confidence of assignment to each class within the selected model.

	prob1	prob2
Typical	0.96	0.04
High	0.29	0.71

[tbp]

Table A5

Levels of classification confidence for each class within the selected model.

	Typical	High
prob>0.7	93.39	64.18
prob>0.8	91.32	7.46
prob>0.9	86.78	0.00

14a. Is a plot included with the estimated mean trajectories of the final solution?

The trajectories are depicted within Figure 2 of the accompanying manuscript, and in isolation in Figure A2.

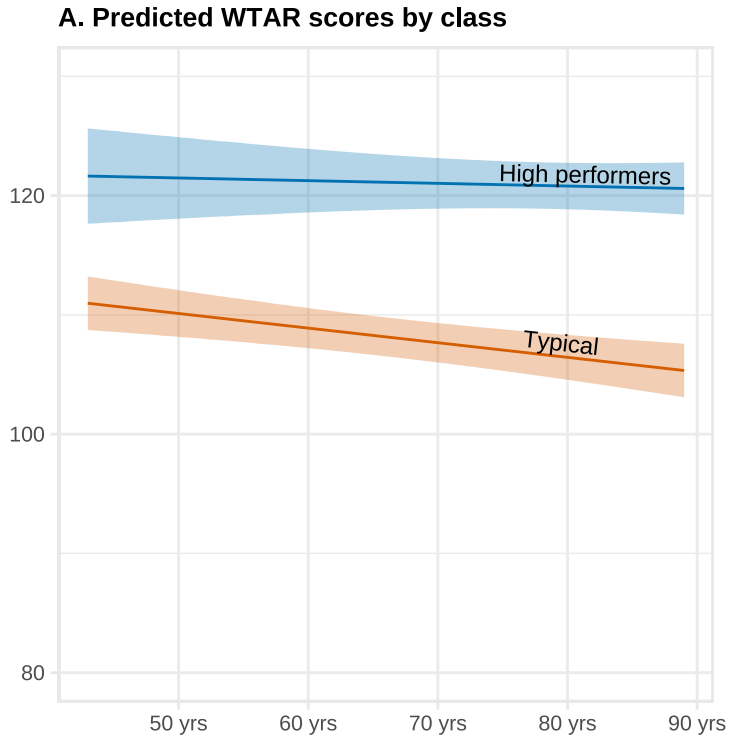

Figure A2. The predicted trajectories of the chosen two-class linear model.

14b. Are plots included with the estimated mean trajectories for each model?

Figure A3 shows the predicted trajectories of all 5 candidate linear and cubic models, as tabulated in Tables A1 and A2.

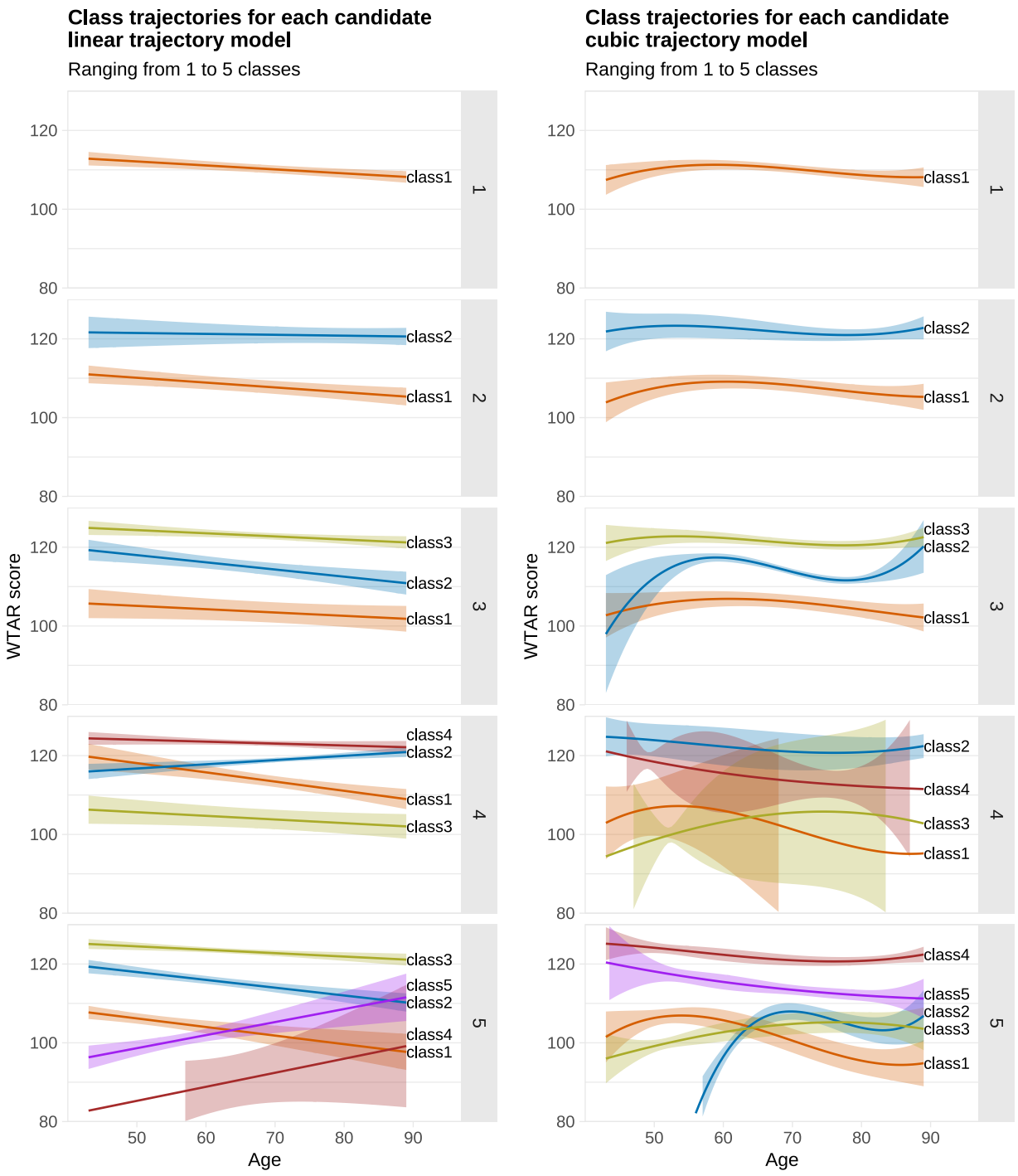

Figure A3. The predicted trajectories of all candidate latent trajectory models (the selected one was the two-class linear trajectory model, as shown in Figure A2). The percentage of participants assigned to each class in each model can be found in Tables A1 and A2.

14c. Is a plot included of the combination of estimated means of the final model and the observed individual trajectories split out for each latent class?

This is shown in Figure 2 of the accompanying manuscript.

15.Are characteristics of the final class solution numerically described (i.e., means, SD/SE, n, CI, etc.)?

Yes, see the Results section of the accompanying manuscript.

16. Are the syntax files available (either in the appendix, supplementary materials, or from the authors)?

Yes, they are publicly available at github.com/nzbri/wtar-trajectory.