 doi:10.20944/preprints202102.0502.v1

Correlation between the Bilingual Status and the Onset Age of AD and MCI Subjects: Evidence from the ADNI dataset

Preliminary version

Jason Li[#, 1], Yang Han[#, 1], Jacqueline CK Lam[#, *, 1], Victor OK Li[#, *, 1], Stephen Matthews[2], Lawrence YL Cheung[3], Virginia Yip[3], Jocelyn Downey[1], Danny Chan[4], Illana Gozes[5], for the Alzheimer's Disease Neuroimaging Initiative[6]

[1]Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong

[2]Department of Linguistics, The University of Hong Kong, Pokfulam Road, Hong Kong

[3]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

[4]School of Biomedical Sciences, The University of Hong Kong, Pokfulam Road, Hong Kong

[5]Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Adams Super Center for Brain Studies and Sagol School of Neuroscience, Tel Aviv University, Israel

[6]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[#]These authors have equal contributions.

*Corresponding authors. Address: Department of Electrical and Electronic Engineering, The University of Hong Kong, Pok Fu Lam, Hong Kong. Email: jcklam@eee.hku.hk; vli@eee.hku.hk

Abstract

Background: This paper investigates the statistical relationship between bilingualism and the Onset Age (OA) of AD and MCI across a clinical sample, consisting of 580 Alzheimer's Disease (AD) subjects and 1264 Mild Cognitive Impairment (MCI) subjects, via a statistical analysis conducted on the sample retrieved from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset.

Method: To investigate whether bilingualism has any correlation with the OAs of AD or MCI subjects, our study leverages the full potential of the ADNI dataset, a dataset that covers both the OA and the bilingualism status of both the AD and MCI subjects. Prior to performing any

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

meaningful statistical analysis, a regression model and a probabilistic model were developed in parallel to fill in the missing OA and bilingualism values. A simple least-square regression model that consists of an independent variable of registered age for Mini-Mental State Examination (MMSE) score was used to estimate the OA of the AD and MCI subjects in the ADNI dataset. After filling in the missing OA values, the number of subjects relevant for the statistical analysis increased from 816 (AD: 371, MCI: 445) to 1844 (AD: 580, MCI: 1264), which greatly enlarged the representation of the AD and MCI sample in the ADNI population. With increased sample size, a novel probabilistic classification model was introduced to infer an ADNI subject's bilingualism when relevant demographic information and deterministic outcome were not readily available from the ADNI dataset. The weighted average OA for the bilinguals and the monolinguals was then computed, where the weights for the probabilistic labels were assigned based on the percentage of bilingualism in the general US population. Finally, a statistical analysis was performed to test whether any statistically significant correlation exists between the OA and the bilingualism of the AD and MCI subjects within the ADNI dataset.

Findings: Our preliminary study demonstrates no significant statistical difference between the OA of the bilinguals and the monolinguals within the ADNI dataset. Thus, the monolingual speakers within the ADNI dataset do not statistically manifest earlier onset, as compared to the bilingual speakers, which is slightly inconsistent with some earlier statistical findings that bilingual speakers enjoy certain distinctive advantages, such as late onset of AD, as compared to monolingual counterparts.

Keywords: Alzheimer's Disease, Onset Age, Bilingualism, Cognitive Reserve, Dementia, Mild Cognitive Impairment, ADNI database

Significance Statement: This paper seeks to overcome a limitation within previous studies, namely, small sample size. By making appropriate assumptions, data were interrogated from 580 AD and 1264 MCI subjects, to investigate multiple AD and linguistic data from the ADNI dataset. This study also provides a way to manage non-deterministic linguistic outcomes to facilitate more rigorous statistical analysis. Most importantly, our preliminary statistical study that investigates the correlation between OA of AD/MCI and bilingualism based on, ADNI, a large clinical dataset shows no conclusive distinctive advantage for bilinguals over monolinguals in terms of delayed AD/MCI onset. Thus, more in-depth investigation might be

 doi:10.20944/preprints202102.0502.v1

needed to identify the difference in OA of AD/MCI for bilingual speakers and monolingual speakers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

1. Introduction

Dementia is a general description for a set of symptoms associated with the deterioration of cognitive abilities such as an individual's episodic memory, verbal skills, reasoning etc., that can gradually affect one's ability to perform daily activities. According to the WHO, it is estimated that around 50 million people are affected by dementia worldwide, with this figure increasing at a rate of 10 million new cases per year, and creating a substantial economic burden of almost a trillion dollars (WHO, 2020).

Alzheimer's Disease (AD) is an irreversible and progressive neurodegenerative disease which accounts for up to 70% of cases of dementia, and, therefore represents a high priority for the development of effective interventional strategies. Such approaches that aim to prevent neurodegenerative disorders such as AD and Mild Cognitive Impairment (MCI) can be generally divided into primary, secondary and tertiary prevention, according to the different stages of disease development (Fratiglioni et al., 2007). This study will focus on primary prevention by identifying protective factors that may decrease or delay the development of AD. Several studies (Alladi et al., 2013; Bak et al., 2014; Bialystok et al., 2014) provided evidence that bilingualism may be a contributing factor that helps to defer the onset of symptoms of AD. This study makes use of evidence from one of the largest AD datasets, the ADNI dataset, to investigate the claim that bilingualism possesses preventive effects against neurodegenerative disorders and delays the OA of AD and MCI.

2. Related Work

2.1 Cognitive Reserve

The concept of cognitive reserve was proposed by Stern (2009) to explain the discrepancy between the degree of brain pathology and clinical manifestations. Previous studies reported that 25% of elderly who performed normally during neuropsychological tests were found to meet the full pathologic criteria for AD, indicating that some people can cope better than others despite similar degrees of brain damage. Cognitive reserve is believed to provide a level of resistance to neurological damage, possibly as the result of increased synaptic plasticity, compensatory use of alternate brain areas, or enriched brain vasculature (Fratiglioni et al., 2004). This prompts us to find contributing factors to cognitive reserve that may help postpone the onset of symptoms of AD and MCI.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

2.2 Defining Bilingualism

Previous studies argued that various environmental factors may affect the onset age of AD. Fratiglioni et al. (2004) suggested that an active and socially integrated lifestyle in late life may impose protective effects against AD. Scarmeas et al. (2001) found that participation in leisure activities decreases the likelihood of incident dementia and may provide a level of cognitive reserve that delays the onset of symptoms of dementing diseases. Fratiglioni et al. (2007) further investigated potential risk factors for AD throughout a lifespan, and argued that that socio-economic factors, including educational level and occupation, as well as life-habits may also affect the risk of AD.

In addition to the features discussed above, there is growing evidence that bilingualism may also defer the onset of the symptoms of AD and MCI. Bialystok et al. (2007) investigated the impact of lifelong bilingualism on the preservation of cognitive functioning and its ability to delay the clinical manifestation of dementia in old age. The study revealed that bilingual individuals exhibited signs of dementia 4 years later than monolinguals, while the rate of decline in Mini-Mental State (MMSE) test scores over the 4-year post-diagnosis period remained the same, indicating a shift in OA with little difference in the rate of development. Craik et al. (2010) looked further into the effect of bilingualism on onset of AD by eliminating the effects of other confounding factors including education, occupational status and immigration, to reveal that bilingualism does indeed contribute to cognitive reserve and delays the onset of AD. Schweizer et al. (2012) adopted a different approach, analysing linear measurements of brain atrophy from computed tomography (CT) scans of bilingual and monolingual individuals with probable AD with matched levels of cognitive performance and education. Their study indicated that bilingual AD patients possess significantly higher levels of cerebral atrophy in areas closely related to AD, supporting the claim that bilingualism leads to increased cognitive reserve. Evidence from studies on cerebral glucose metabolism in MCI and AD also support a role for bilingualism in increasing cognitive reserve, reporting that bilingual patients have more severe brain changes than monolinguals when adjusting for severity of cognitive impairment (Kowoll et al., 2016).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

3. Methodology

After reviewing previous studies investigating the effect of bilingualism on AD, we have identified the key factors to be taken into account in the statistical analysis (see Table 1). The commonly used confounding variables include education, MMSE score, occupation, lifestyle, ethnicity, race, language proficiency, etc. With most of the clinical studies mainly carried out on relatively small sample sizes (n<200), our study intends to leverage extensive AD databases and see if the pattern holds true over larger sample sizes. In particular, we utilized the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, which is one of the most comprehensive, precise and readily available AD dataset currently (Mueller, 2005). Nevertheless, the ADNI dataset has no label for bilingualism as well as explicit information about subjects' proficiency in their primary and secondary language. Thus, we propose a methodology to create a probabilistic bilingual classification that consists of three steps (see Figure 1). First, we selected AD and MCI subjects from the ADNI dataset and extracted their information relevant to our study, including OA and demographic information. Second, we filled in the missing OA values with a regression model, and assigned a probability classification to subjects not explicitly determined to be bilinguals/monolinguals. Third, based on the pre-processed data, we examined whether bilingualism has any correlation with the OAs of AD or MCI subjects.

Table 1. Confounding factors used in the previous bilingualism and AD studies

Study	**No. of Subjects**	**Confounding Factors**	**Models**
Bialystok et al. (2007)	AD (n=184): M = 91, B = 93	Immigration, age at the first appointment, education, MMSE score at the first appointment, occupation status	Two-way ANOVA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

Gollan et al. (2011)	AD (n=44): B = 44	education, degree of bilingualism	Pearson bivariate correlations
Bialystok et al. (2014)	MCI (n=74): M = 38, B = 36 AD (n=75): M = 35, B = 40	Immigration, education, lifestyles (diet, alcohol, smoking, physical and social activities)	Two-way ANOVA, Partial correlation coefficients
Woumans et al. (2015)	AD (n=134):M = 69, B = 65	gender (factor), education (in years), occupation (three levels), MMSE at diagnosis	Linear regression model
Calabria et al. (2020)	MCI (n=135): M = 38, B = 36 AD (n=68): M = 35, B = 40 CN (n=63): M = 35, B = 40	bilingualism composite factor*, cognitive decline composite factor, CRIq score	Regression models, Partial correlation coefficients

* The bilingualism composite factor was calculated by performing a PCA using variables: years of language exposure to both languages for speaking; self-rating of language proficiency in both languages for speaking, comprehension, writing, and reading; the percentage of language usage; and frequency of language switching.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

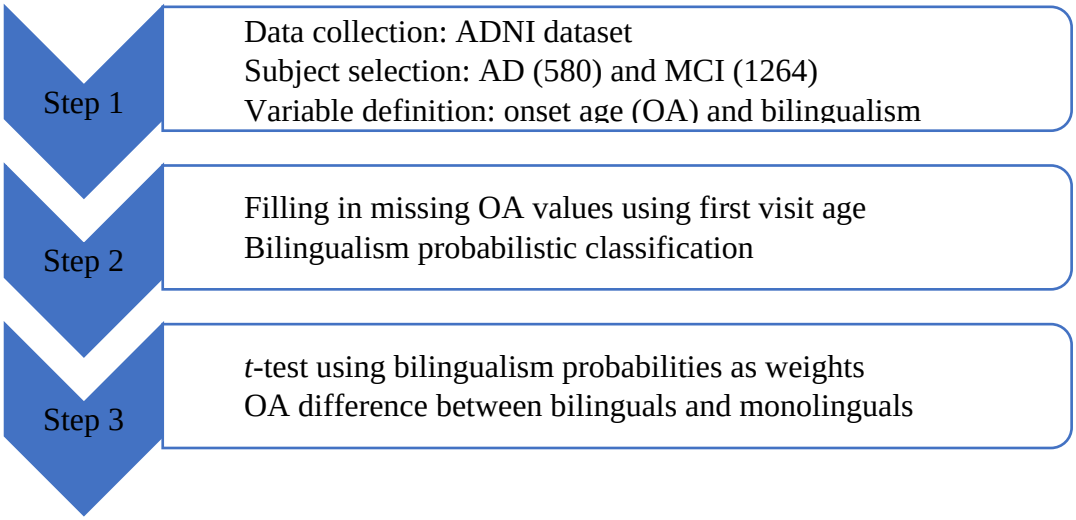


Figure 1. Overall methodology

3.1 ADNI Dataset

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org. The ADNI datasets were downloaded from the ADNI server (http://adni.loni.ucla.edu/) on 17 January 2021.

3.2 Subject Selection and Variable Definition

We have selected a total of 1844 (AD: 580, MCI: 1264) subjects from the ADNI dataset. We then extract the relevant features from the dataset for statistical analysis (Table 2).

Table 2. Relevant features in the ADNI dataset that can be used as control variables

Feature	Description	Type	Note
MMSE	Mini Mental State Exam	Number	Multiple MMSE scores are available starting from the first visit

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

ADNI-LAN	A composite score for language-related task performance (using the tested language)	Number	Language tasks include reading, writing, repeating sentences, etc.
PTTLANG	Language to be used for testing the Participant	Number	1=English 2=Spanish
PTPLANG	Participant's Primary Language	Number	1=English 2=Spanish 3=Other (specify)
PTEDUCAT	Participant Education	Number	0...20
PTWORKHS	Does the participant have a work history sufficient to exclude mental retardation?	Boolean	1 = yes 0 = no
PTNOTRT	Participant Retired?	Boolean	1 = yes 0 = no
PTRTYR	Retirement Date	Date	
PTMCIBEG	Year of onset of Mild Cognitive Impairment symptoms (best estimate)	Number	

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

PTADBEG	Year of onset of Alzheimer's disease symptoms (best estimate)	Number	
PTETHCAT	Ethnic Category	Number	1=Hispanic or Latino 2=Not Hispanic or Latino 3=Unknown
PTRACCAT	Racial Categories	Number	1=American Indian or Alaskan Native; 2=Asian; 3=Native Hawaiian or Other Pacific Islander; 4=Black or African American; 5=White; 6=More than one race; 7=Unknown
PTDOBYY	Participant Year of Birth	Number	

After extraction of the data, we perform the following data treatment:

Onset age of AD and MCI (dependent variable): By subtracting a participant's year of birth (PTDOBYY) from the participant's year of onset for MCI (PTMCIBEG) and for AD (PTADBEG), we obtain the OA for AD and MCI, respectively. According to the ADNI dataset, the OA of an AD/MCI subject is defined as the reported age when a subject believes his AD/MCI symptoms have begun (see Equations (1) and (2)), so the diagnostic age definition is used.

$$OA\,(AD) = PTADBEG - PTDOY \quad (1)$$
$$OA\,(MCI) = PTMCIBEG - PTDO \quad (2)$$

Bilingualism (independent variable): According to Hamers et al. (2000), bilingualism consists of a wide range of definitions ranging from a native-like competence in both languages

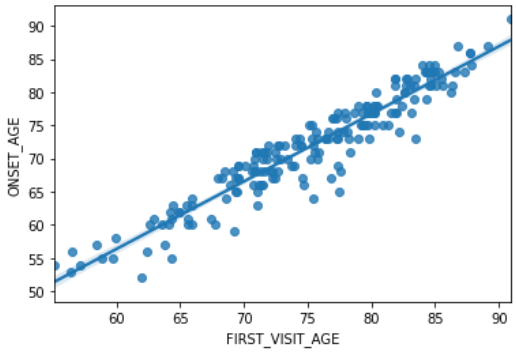
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

to a minimal proficiency in a second language. In order to prevent any confusion, this study defines bilingualism as mastering the second language with minimal proficiency. The ADNI dataset does not indicate the bilingual status of individuals. However, it includes information on primary language and tested language (the language that the subject had been tested when admitted to the ADNI clinical trial). We take subjects who speak a primary language different from his/her tested language as bilinguals. For the remaining subjects, we adopt a probabilistic classification approach to determine the subject's bilingualism (see Section 3.3.2 for more details).

3.3 Filling in Missing OA and Bilingualism Label

3.3.1 Filling in Missing OA Values Using First Visit Age

Since most of the OA values of the MCI and AD subjects in the ADNI dataset are missing, a least-square regression model is used to fill in the missing OA values. A linear regression model with a total of 1844 observations is fitted using the age of an AD/MCI subject when the subject visited/tested for the first time during the ADNI study (also referred to as first visit age) to predict the subject's OA. Figure 2 shows the results of the fitted regression model for estimating the OA based on the first visit age. From the model, first visit age is a significant predictor of the OA with $p< 0.001$. Moreover, $R^2>0.999$ also indicates that the regression model is able to explain most of the variance in OA by using the first visit age only.



(a)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

OLS Regression Results

Dep. Variable:	ONSET_AGE	**R-squared (uncentered):**	0.999
Model:	OLS	**Adj. R-squared (uncentered):**	0.999
Method:	Least Squares	**F-statistic:**	1.550e+05
Date:	Thu, 21 Jan 2021	**Prob (F-statistic):**	6.89e-275
Time:	16:47:33	**Log-Likelihood:**	-439.90
No. Observations:	188	**AIC:**	881.8
Df Residuals:	187	**BIC:**	885.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
FIRST_VISIT_AGE	0.9561	0.002	393.700	0.000	0.951	0.961

Omnibus:	32.608	**Durbin-Watson:**	2.113
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	44.627
Skew:	-1.042	**Prob(JB):**	2.04e-10
Kurtosis:	4.165	**Cond. No.**	1.00

(b)

Figure 2. (a) Scatter plot of OA and first visit age and (b) the results of using a regression model to recover OA based on the age at first visit

3.3.2 Bilingualism Probabilistic Classification

A major obstacle in delineating the statistical relationship between the OA and bilingualism of AD and MCI subjects using the ADNI dataset is that there is no explicit label describing the level of bilingualism of the subjects. In order to tackle this, we propose a rule-based method which assigns probabilistic labels to participants based on four other available features: (1) Language to be used for testing the Participant (PTTLANG), (2) Participant's Primary Language (PTPLANG), (3) Participant's Ethnic Category (PTETHCAT) and (4) Participant's Racial Category (PTRACCAT).

In order to assign a probabilistic label to the participant's bilingualism based on his or her ethnicity, we will use the ethnicity and language characteristics of the US population from US census data as our baseline data (United States Census Bureau, 2019). Table 3 shows the statistics describing the language use at home of the population in the US based on their ethnicity.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

Table 3. Language spoken at home by nativity from the ACS 5-year estimates (2015-2019)

Ethnicity	**Total**	**Hispanic or Latino**	**Non-Hispanic**
Total	304930125	53372815	251557310
Native	261219994	33899836	227320158
Native - Speak only English	231864766	14023383	217841383
Native - Speak another language	29355228	19876453	9478775
Native - Speak another language (Speak English very well)	24593283	16504182	8089101
Foreign born	43710131	19472979	24237152
Foreign born - Speak only English	7117586	979093	6138493
Foreign born - Speak another language	36592545	18493886	18098659
Foreign born - Speak another language (Speak English very well)	15739125	6018516	9720609
% of US Population	93.9%	16.4%	77.5%
% of Foreign-born	14.3%	36.5%	9.6%
% of Speak only English	78.4%	28.1%	89.0%
% of Speak another language*	21.6%	71.9%	11.0%
% of Speaking another language (Speak English very well)	13.2%	42.2%	8.0%
* The percentage of people who can not only speak English but also speak another language (such as Spanish) at home.			

Two sets of experiments are performed to simulate two different scenarios, Scenario A and Scenario B, as illustrated in Table 4, to observe whether the weighted average OA will be consistent. After going through one of the two scenarios in Table 4, each subject was assigned either one (probability of zero or one) or two (probabilities for being bilingual and monolingual) probabilistic labels. For example, if a subject is Hispanic, two entries will be created in the dataset, one with a bilingual label and one with a monolingual label, with probabilities of 0.719 and 0.281, respectively. If a subject's primary and tested language is different, then we determine this subject is bilingual with probability one, and so only one entry

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

will be created. Finally, the average AD/MCI OA weighted by probability can be computed for bilingual entries and monolingual entries separately.

Table 4. Assigning probabilistic labels to participants' bilingualism (B: bilingual; M: monolingual)

Scenario A	**B**	**M**	**Prob.**
● If the subject's primary language is different from his/her testing language, then it is replaced with one entry, with a bilingual label and probability one	1	0	100%
● Else if (1) the subject's primary language and testing language are both English and (2) his/her race is Black American or White American, it is replaced with a monolingual label and probability one	0	1	100%
● Else, a subject is replaced with two entries: ○ One with the bilingual label and the bilingual probability conditioning on his/her ethnicity ○ Another with the monolingual label and the monolingual probability conditioning on his/her ethnicity ○ Using information based on the 2019 US census data, specifically the percentage of the US population speaking English only.	**Non-Hispanics**		
	1	0	11%
	0	1	89%
	Hispanics		
	1	0	71.9%
	0	1	28.1%
Scenario B	**B**	**M**	**Prob.**
● If the subject's primary language is different from his/her testing language, then it is replaced with one entry, with a bilingual label and probability one	1	0	100%
● Else, a subject is replaced with two entries: ○ One with the bilingual label and the bilingual probability conditioning on his/her ethnicity ○ Another with the monolingual label and the monolingual probability conditioning on his/her ethnicity ○ Using information based on the 2019 US census data, specifically the percentage of the US population speaking English only.	**Non-Hispanics**		
	1	0	11%
	0	1	89%
	Hispanics		
	1	0	71.9%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

	0	1	28.1%

3.4 Statistical Analysis on the Correlation between OA and Bilingualism

We tested the statistical correlation between OA and bilingualism using a two-sided *t*-test. More specifically, we adopted the Welch's *t*-test, which does not assume that the two groups have equal variance. The Welch's *t*-test evaluated the mean difference between the bilingual and the monolingual groups (probabilistic entries), using their corresponding probabilities as the weights. The null hypothesis is that there is no OA difference between the bilinguals and the monolinguals. A *p*-value greater than 0.05 is considered not statistically significant, indicating strong evidence for the null hypothesis.

4. Results

The statistical results for the aforementioned two scenarios with *p*-values are listed in Table 5. In general, the *p*-values range from 0.43 to 0.89 and from 0.15 to 0.93 for AD subjects and MCI subjects, respectively. No OA difference can be observed at $p<0.05$ or $p<0.01$, across all scenarios and all population groups (including Hispancs and Non-Hispanics).

Table 5. Average OA of AD and MCI subjects

Scenario A		
All		
Category	**AD (580 subjects)**	**MCI (1264 subjects)**
Average OA (Bilingual)	70.4 (9 for sure)	67.5 (27 for sure)
Average OA (Monolingual)	71.8 (549 for sure)	69.6 (1194 for sure)
p-value	0.49	0.15
Hispanic		
Category (no. of subjects)	**AD (18 subjects)**	**MCI (49 subjects)**
Average OA (Bilingual)	71.2 (3 for sure)	65.4 (12 for sure)
Average OA (Monolingual)	68.8 (6 for sure)	67.2 (22 for sure)
p-value	0.57	0.47

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

Non-Hispanic		
Category	**AD (562 subjects)**	**MCI (1215 subjects)**
Average OA (Bilingual)	69.3 (6 for sure)	70.1 (15 for sure)
Average OA (Monolingual)	71.9 (543 for sure)	69.6 (1172 for sure)
p-value	0.43	0.84
Note: The number of entries is shown in parenthesis (they are not equal to the number of subjects due to probabilistic assignment).		

Scenario B		
All		
Category	**AD (580 subjects)**	**MCI (1264 subjects)**
Average onset age (Bilingual)	71.3 (9 for sure)	69.0 (27 for sure)
Average onset age (Monolingual)	71.9 (0 for sure)	69.6 (0 for sure)
p-value	0.60	0.34
Hispanic		
Category (no. of subjects)	**AD (18 subjects)**	**MCI (49 subjects)**
Average onset age (Bilingual)	69.8 (3 for sure)	66.2 (12 for sure)
Average onset age (Monolingual)	70.6 (0 for sure)	66.9 (0 for sure)
p-value	0.89	0.84
Non-Hispanic		
Category	**AD (562 subjects)**	**MCI (1215 subjects)**
Average onset age (Bilingual)	71.6 (6 for sure)	69.7 (15 for sure)
Average onset age (Monolingual)	71.9 (0 for sure)	69.6 (0 for sure)
p-value	0.82	0.93
Note: The number of entries is shown in parenthesis (they are not equal to the number of subjects due to probabilistic assignment).		

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

5. Discussions and Future Work

Given that at this stage, no OA difference can be observed at the significance level of at $p<0.05$ or $p<0.01$, for the AD/MCI subjects in the ADNI dataset, no definitive conclusions can be reached on whether bilingualism can postpone the OA of AD/MCI. With our study based on a statistically rigorous methodology and a sufficiently large AD/MCI dataset, we suggest that it might be worthwhile to re-examine the statistical relationship between OA and bilingualism for MCI and AD subjects. Nevertheless, some challenges remain to be overcome in the future study:

Diagnostic age vs. manifestation age: With our method of estimating OA with first-visit age, we are equating OA with diagnostic age, which is the objective estimate by professional clinical classification, rather than the manifestation age, the subjective estimate by patients and their families. Given that OA may be a subjective definition, further studies should include the diagnostic age as another dependent variable in the statistical analysis, assuming that the diagnosis provided in the ADNI study is reliable.

Bilingualism classification for the elderly population: In this study, there is a mismatch between the ADNI population and the US census population. More specifically, the census data across all age groups was used to estimate the bilingual ratio among Hispanics and non-Hispanics. In the future, a more detailed breakdown of the US census data, including age and ethnicity data, could be used to derive the bilingualism probability for the elderly population.

The degree of bilingualism: In this study, the degree of bilingualism was not taken into account. The US census data has provided two labels to determine the degree of bilingualism (for people who speak another language spoken at home such as Spanish): speaking English very well and speaking English less than very well. In the future, we will evaluate whether or not the effects of bilingualism on OA are different with regard to the three levels of bilingualism (i.e., English only, speaking another language and speaking English less than very well, and speaking another language and speaking English very well).

The effects of bilingualism on certain cognitive tasks: In this study, we only investigated the relationship between bilingualism and OA. However, bilinguals may only have a certain advantage over certain tasks. The ADNI dataset has provided four composite measures, each representing a different aspect of cognitive ability: ADNI-EF (executive functioning), ADNI-MEM (memory), ADNI-LAN (Language), and ADNI-VS (visuospatial

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2021 doi:10.20944/preprints202102.0502.v1

functioning). In the future, we will examine the effects of bilingualism on ADNI-LAN and other cognitive abilities.

Confounding bias: In this study, the statistical association between OA and bilingualism may be biased due to uncontrolled confounding factors. Further studies that account for the important confounding factors, such as education, occupation, lifestyle, physical activities, and diet etc., are needed. For example, analysis of covariance (ANCOVA) can be performed by controlling for occupation and education. These factors can be converted into quantitative indicators, e.g., using a 5-category scale according to the International Standard Classification of Occupation for occupation history (International Labor Office, 2012).

6. Conclusions and Policy Recommendations

Our study explored the relationship between the OA of AD/MCI patients and their level of bilingualism using ADNI, a comprehensive AD/MCI database that consists of 1844 AD/MCI subjects. We proposed a novel methodology to predict missing OA values and level of bilingualism with a linear regression model and probabilistic classification model respectively, providing an alternative to manage non-deterministic linguistic outcomes and facilitate more rigorous statistical analysis. Based on our statistical model, the difference in OA between monolingual and bilingual AD/MCI patients is insignificant (p>0.05). This suggests that more in-depth studies to investigate the effects of other compounding factors such as education and occupation levels and extensive linguistic data collection from AD/MCI patients are required.

While ADNI is already one of the most extensive and comprehensive AD/MCI databases available worldwide, the fact that the subjects' level of bilingualism are yet to be provided by ADNI, suggests that there is a general lack of awareness that linguistic features can be good indicators or markers of AD . Thus, there is a strong need for health policy-makers and research communities worldwide to collect more extensive linguistic data from the AD/MCI population, as well as more rigorous studies to be conducted internationally and locally, to facilitate linguistic-based investigation of AD and evidence-based health policymaking for the vulnerable AD communities.

7. Acknowledgement

 doi:10.20944/preprints202102.0502.v1

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. We also sincerely acknowledge the contribution of Mr. Kenyon Chow, PhD student of HKU-Cambridge AI-WiSe Research Platform, EEE Dept., HKU, for his careful revision and editing of the current manuscript. All errors remain ours.

8. References

Alladi, S., Bak, T. H., Duggirala, V., Surampudi, B., Shailaja, M., Shukla, A. K., ... & Kaul, S. (2013). Bilingualism delays age at onset of dementia, independent of education and immigration status. *Neurology*, *81*(22), 1938-1944.

Anderson, J. A., Hawrylewicz, K., & Grundy, J. G. (2020). Does bilingualism protect against dementia? A meta-analysis. *Psychonomic Bulletin & Review*.

Bak, T. H., Nissan, J. J., Allerhand, M. M., & Deary, I. J. (2014). Does bilingualism influence cognitive aging?. *Annals of neurology*, *75*(6), 959-963.

Bialystok, E., Craik, F. I., & Freedman, M. (2007). Bilingualism as a protection against the onset of symptoms of dementia. *Neuropsychologia*, *45*(2), 459-464.

Bialystok, E., Craik, F. I., Binns, M. A., Ossher, L., & Freedman, M. (2014). Effects of bilingualism on the age of onset and progression of MCI and AD: Evidence from executive function tests. *Neuropsychology*, *28*(2), 290.

Craik, F. I., Bialystok, E., & Freedman, M. (2010). Delaying the onset of Alzheimer disease: Bilingualism as a form of cognitive reserve. *Neurology*, *75*(19), 1726-1729.

de Leon, J., Grasso, S. M., Welch, A., Miller, Z., Shwe, W., Rabinovici, G. D., ... & Gorno-Tempini, M. L. (2020). Effects of bilingualism on age at onset in two clinical Alzheimer's disease variants. *Alzheimer's & Dementia*.

Esiri, M. M., Matthews, F., Brayne, C., Ince, P. G., Matthews, F. E., Xuereb, J. H., ... & Lowe, J. (2001). Pathological correlates of late-onset dementia in a multicentre, community-based population in England and Wales. *Lancet*, *357*(9251).

Fratiglioni, L., Paillard-Borg, S., & Winblad, B. (2004). An active and socially integrated lifestyle in late life might protect against dementia. *The Lancet Neurology*, *3*(6), 343-353.

Fratiglioni, L., Winblad, B., & von Strauss, E. (2007). Prevention of Alzheimer's disease and dementia. Major findings from the Kungsholmen Project. *Physiology & behavior*, *92*(1-2), 98-104

Gold, B. T. (2015). Lifelong bilingualism and neural reserve against Alzheimer's disease: A review of findings and potential mechanisms. *Behavioural Brain Research*, *281*, 9-15.

 doi:10.20944/preprints202102.0502.v1

Gollan, T. H., Salmon, D. P., Montoya, R. I., & Galasko, D. R. (2011). Degree of bilingualism predicts age of diagnosis of Alzheimer's disease in low-education but not in highly educated Hispanics. *Neuropsychologia*, *49*(14), 3826-3830.

Government of Canada. (2020). Find your NOC. Retrieved from https://www.canada.ca/en/immigration-refugees-citizenship/services/immigrate-canada/express-entry/eligibility/find-national-occupation-code.html

Giles, E., Patterson, K., & Hodges, J. R. (1996). Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer's type: missing information. *Aphasiology*, *10*(4), 395-408.

Hamers, J. F., Blanc, M., & Blanc, M. H. (2000). *Bilinguality and bilingualism*. Cambridge University Press.

International Labor Office. (2012). International Standard Classification of Occupations: Structure, group definitions and correspondence tables.Retrieved from https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_172572.pdf

Klimova, B., Valis, M., & Kuca, K. (2017). Bilingualism as a strategy to delay the onset of Alzheimer's disease. *Clinical Interventions in Aging*, *12*, 1731.

Kowoll, M. E., Degen, C., Gorenc, L., Küntzelmann, A., Fellhauer, I., Giesel, F., ... & Schröder, J. (2016). Bilingualism as a contributor to cognitive reserve? Evidence from cerebral glucose metabolism in mild cognitive impairment and Alzheimer's disease. *Frontiers in psychiatry*, *7*, 62.

Krogstad, J., & Gonzalez-Barrera, A. (2015). A majority of English-speaking Hispanics in the U.S. are bilingual. Retrieved from https://www.pewresearch.org/fact-tank/2015/03/24/a-majority-of-english-speaking-hispanics-in-the-u-s-are-bilingual/

Lombardi, G., Polito, C., Berti, V., Bagnoli, S., Nacmias, B., Pupi, A., & Sorbi, S. (2018). Contribution of bilingualism to cognitive reserve of an Italian literature professor at high risk for Alzheimer's disease. *Journal of Alzheimer's Disease*, *66*(4), 1389-1395.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., ... & Beckett, L. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, *15*(4), 869-877.

Paulavicius, A. M., Mizzaci, C. C., Tavares, D. R., Rocha, A. P., Civile, V. T., Schultz, R. R., ... & Trevisani, V. F. (2020). Bilingualism for delaying the onset of Alzheimer's disease: a systematic review and meta-analysis. *European Geriatric Medicine*, 1-8.

Scarmeas, N., Levy, G., Tang, M. X., Manly, J., & Stern, Y. (2001). Influence of leisure activity on the incidence of Alzheimer's disease. *Neurology*, *57*(12), 2236-2242.

Scarmeas, N., Albert, S. M., Manly, J. J., & Stern, Y. (2006). Education and rates of cognitive decline in incident Alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *77*(3), 308-316.

Schweizer, T. A., Ware, J., Fischer, C. E., Craik, F. I., & Bialystok, E. (2012). Bilingualism as a contributor to cognitive reserve: Evidence from brain atrophy in Alzheimer's disease. *cortex*, *48*(8), 991-996.

Stern, Y. (2009). Cognitive reserve. *Neuropsychologia*, *47*(10), 2015-2028.

U.S. National Library of Medicine. (2005). ADNI: Alzheimer's Disease Neuroimaging Initiative. Retrieved from https://clinicaltrials.gov/ct2/show/NCT00106899.

United States Census Bureau. (2019). ACS Demographic and Housing Estimates. Retrieved from https://data.census.gov/cedsci/table?d=ACS%205-Year%20Estimates%20Data%20Profiles&tid=ACSDP5Y2019.DP05.

 doi:10.20944/preprints202102.0502.v1

Valenzuela, M. J., & Sachdev, P. (2006). Brain reserve and cognitive decline: a non-parametric systematic review. *Psychological medicine*, *36*(8), 1065-1073.

Voits, T., Pliatsikas, C., Robson, H., & Rothman, J. (2020). Beyond Alzheimer's disease: Can bilingualism be a more generalized protective factor in neurodegeneration?. *Neuropsychologia*, 107593.

Weimer, D. L., & Sager, M. A. (2009). Early identification and treatment of Alzheimer's disease: social and fiscal outcomes. *Alzheimer's & Dementia*, *5*(3), 215-226.

Williams, J. A., Weakley, A., Cook, D. J., & Schmitter-Edgecombe, M. (2013, June). Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. In *Workshops at the twenty-seventh AAAI conference on artificial intelligence.*

World Health Organization (2020). Dementia. Retrieved from https://www.who.int/news-room/fact-sheets/detail/dementia.

Woumans, E. V. Y., Santens, P., Sieben, A., Versijpt, J. A. N., Stevens, M., & Duyck, W. (2015). Bilingualism delays clinical manifestation of Alzheimer's disease. *Bilingualism: Language and Cognition*, *18*(3), 568-574.

Zhang, Y., Londos, E., Minthon, L., Wattmo, C., Liu, H., Aspelin, P., & Wahlund, L. O. (2008). Usefulness of computed tomography linear measurements in diagnosing Alzheimer's disease. *Acta radiologica*, *49*(1), 91-97.