Personalized genetic assessment of age-associated Alzheimer's disease risk

Rahul S. Desikan, MD, PhD[1*#]; Chun Chieh Fan, MD[2*]; Yunpeng Wang, PhD[3,4]; Andrew J. Schork, MS[2]; Howard J. Cabral, PhD[5]; L. Adrienne Cupples, PhD[5]; Wesley K. Thompson, PhD[6]; Lilah Besser, MSPH[7]; Walter A. Kukull, PhD[7]; Dominic Holland, PhD[3]; Chi-Hua Chen, PhD[8]; James B. Brewer, MD, PhD[3,8,19]; David S. Karow, MD, PhD[8]; Karolina Kauppi, PhD[8]; Aree Witoelar, PhD[4]; Celeste M. Karch, PhD[9]; Luke W. Bonham, BS[10]; Jennifer S. Yokoyama, PhD[10]; Howard J. Rosen, MD[10]; Bruce L. Miller, MD[10]; William P. Dillon, MD[1]; David M. Wilson, MD, PhD[1]; Christopher P. Hess, MD, PhD[1]; Margaret Pericak-Vance, PhD[11]; Jonathan L. Haines, PhD[12]; Lindsay A. Farrer, PhD[13]; Richard Mayeux, MD[14]; John Hardy, PhD[15]; Alison M. Goate, PhD[16]; Bradley T. Hyman, MD, PhD[17]; Gerard D. Schellenberg, PhD[18]; Linda K. McEvoy, PhD[8]; Ole A. Andreassen, MD, PhD[4#]; Anders M. Dale, PhD[2,3,8#] for the ADNI and ADGC investigators

*Contributed equally

[1]Neuroradiology Section, Department of Radiology and Biomedical Imaging, University of California, San Francisco, San Francisco, CA USA
Departments of [2]Cognitive Sciences and [3]Neurosciences, University of California, San Diego, La Jolla, CA, USA
[4]NORMENT; Institute of Clinical Medicine, University of Oslo and Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway
[5]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA
[6]Institute for Biological Psychiatry, Sankt Hans Psychiatric Hospital, Roskilde, Denmark
[7]National Alzheimer's Coordinating Center, Department of Epidemiology, University of Washington, Seattle, WA, USA
[8]Department of Radiology, University of California, San Diego, La Jolla, CA, USA
[9]Department of Psychiatry, Washington University, St. Louis, MO, USA
[10]Department of Neurology, University of California, San Francisco, San Francisco, CA, USA
[11]The John P. Hussman Institute for Human Genomics, University of Miami, Miami, Florida, USA
[12]Department of Epidemiology and Biostatistics and Institute for Computational Biology, Case Western University, Cleveland, Ohio, USA
[13]Departments of Medicine (Biomedical Genetics), Neurology, Ophthalmology, Biostatistics, and Epidemiology, Boston University Schools of Medicine and Public Health, Boston, Massachusetts, USA
[14]Department of Neurology, Taub Institute on Alzheimer's Disease and the Aging Brain, and Gertrude H. Sergievsky Center, Columbia University, New York, New York, USA
[15]Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK
[16]Departments of Neuroscience, Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[17]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA
[18]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[19]Shiley-Marcos Alzheimer's Disease Research Center, University of California, San Diego, La Jolla, CA, USA

#Correspondence should be addressed to:

Dr. Rahul S. Desikan
Neuroradiology Section, L-352
University of California, San Francisco
505 Parnassus Avenue
San Francisco, CA, USA 94143
Email: rahul.desikan@ucsf.edu
Phone: (415)-353-1079

Dr. Ole A. Andreassen
KG Jebsen Centre for Psychosis Research
Building 49, Oslo University Hospital, Ullevål
Kirkeveien 166, PO Box 4956 Nydalen
0424 Oslo, Norway
Email: o.a.andreassen@medisin.uio.no
Ph: +47 23 02 73 50 (22 11 78 43 dir)
Fax: +47 23 02 73 33

Dr. Anders M. Dale
Department of Radiology
University of California, San Diego
8950 Villa La Jolla Drive, Suite C101
La Jolla, CA, USA 92037-0841
Emails: amdale@ucsd.edu
Phone: (858)-822-6671
Fax: (858)-534-1078

*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

Manuscript = 3,095 words

ABSTRACT

Importance: Identifying individuals at risk for developing Alzheimer's disease (AD) is of utmost importance. Although genetic studies have identified *APOE* and other AD associated single nucleotide polymorphisms (SNPs), genetic information has not been integrated into an epidemiological framework for personalized risk prediction.

Objective: To develop, replicate and validate a novel polygenic hazard score for predicting age-specific risk for AD.

Setting: Multi-center, multi-cohort genetic and clinical data.

Participants: We assessed genetic data from 17,008 AD patients and 37,154 controls from the International Genetics of Alzheimer's Project (IGAP), and 6,409 AD patients and 9,386 older controls from Phase 1 Alzheimer's Disease Genetics Consortium (ADGC). As independent replication and validation cohorts, we also evaluated genetic, neuroimaging, neuropathologic, CSF and clinical data from ADGC Phase 2, National Institute of Aging Alzheimer's Disease Center (NIA ADC) and Alzheimer's Disease Neuroimaging Initiative (ADNI) (total n = 20,680)

Main Outcome(s) and Measure(s): Use the IGAP cohort to first identify AD associated SNPs (at $p < 10^{-5}$). Next, integrate these AD associated SNPs into a Cox proportional hazards model using ADGC phase 1 genetic data, providing a polygenic hazard score (PHS) for each participant. Combine population based incidence rates, and genotype-derived PHS for each individual to derive estimates of instantaneous risk for developing AD, based on genotype and age. Finally, assess replication and validation of PHS in independent cohorts.

Results: Individuals in the highest PHS quantiles developed AD at a considerably lower age and had the highest yearly AD incidence rate. Among *APOE* ε3/3 individuals, PHS

115 modified expected age of AD onset by more than 10 years between the lowest and
116 highest deciles. In independent cohorts, PHS strongly predicted empirical age of AD
117 onset ($p = 1.1 \times 10^{-26}$), longitudinal progression from normal aging to AD ($p = 1.54 \times 10^{-10}$) and associated with markers of AD neurodegeneration.
119 **Conclusions:** We developed, replicated and validated a clinically usable PHS for
120 quantifying individual differences in age-specific risk of AD. Beyond *APOE*, polygenic
121 architecture plays an important role in modifying AD risk. Precise quantification of AD
122 genetic risk will be useful for early diagnosis and therapeutic strategies.
123

INTRODUCTION

Late onset Alzheimer's disease (AD), the most common form of dementia, places a large emotional and economic burden on patients and society. With increasing health care expenditures among cognitively impaired elderly[1], identifying individuals at risk for developing AD is of utmost importance for potential preventative and therapeutic strategies. Inheritance of the ε4 allele of apolipoprotein E (*APOE*) on chromosome 19q13 is the most significant risk factor for developing late-onset AD.[2] *APOE* ε4 has a dose dependent effect on age of onset, increases AD risk three-fold in heterozygotes and fifteen-fold in homozygotes, and is implicated in 20-25% of patients with AD.[3]

In addition to *APOE*, recent genome-wide association studies (GWAS) have identified numerous AD associated single nucleotide polymorphisms (SNPs), most of which have a small effect on disease risk.[4-5] Although no single polymorphism may be informative clinically, a combination of *APOE* and non-*APOE* SNPs may help identify older individuals at increased risk for AD. Despite the detection of novel AD associated genes, GWAS findings have not yet been incorporated into a genetic epidemiology framework for individualized risk prediction.

Building on a prior approach evaluating GWAS-detected genetic variants for disease prediction[7] and using a survival analysis framework, we tested the feasibility of combining AD associated SNPs and *APOE* status into a continuous measure 'polygenic hazard score' (PHS) for predicting the age-specific risk for developing AD. We assessed replication and validation of the PHS using several independent cohorts.

METHODS

Participant Samples

IGAP: To select AD associated SNPs, we evaluated publicly available AD GWAS summary statistic data (p-values and odds ratios) from the International Genomics of Alzheimer's Disease Project (IGAP Stage 1, for additional details see Supplemental Information and reference 4). We used IGAP Stage 1 data, consisting of 17,008 AD cases and 37,154 controls, for selecting AD associated SNPs (for a description of the AD cases and controls within the IGAP Stage 1 sub-studies, please see Table 1 and reference 4).

ADGC: To develop the survival model for the polygenic hazard scores (PHS), we first evaluated age of onset and raw genotype data from 6,409 patients with clinically diagnosed AD and 9,386 cognitively normal older individuals provided by the Alzheimer's Disease Genetics Consortium (ADGC, Phase 1, a subset of the IGAP dataset), excluding individuals from the National Institute of Aging Alzheimer's Disease Center (NIA ADC) samples and Alzheimer's Disease Neuroimaging Initiative (ADNI). To evaluate replication of PHS, we used an independent sample of 6,984 AD patients and 10,972 cognitively normal older individuals from the ADGC Phase 2 cohort (Table 1). A detailed description of the genotype and phenotype data within the ADGC datasets has been described in detail elsewhere.[7,24] Briefly, the ADGC Phase 1 and 2 datasets consist of multi-center, case-control, prospective, and family-based sub-studies of Caucasian participants with AD occurrence after age 60. Participants with autosomal dominant (*APP*, *PSEN1* and *PSEN2*) mutations were excluded. All participants were genotyped using commercially available high-density SNP microarrays from Illumina or Affymetrix. Clinical diagnosis of AD within the ADGC sub-studies was established using NINCDS/ADRDA criteria for definite, probable or possible AD. [8] For most participants,

171 age of AD onset was obtained from medical records and defined as the age when AD
172 symptoms manifested, as reported by the participant or an informant. For participants
173 lacking age of onset, age at ascertainment was used. Patients with an age-at-onset or age-
174 at-death less than 60 years, and Caucasians of European ancestry were excluded from the
175 analyses. For additional details regarding the ADGC datasets, please see references 7 and
176 24.
177 NIA ADC: To assess longitudinal prediction, we evaluated an ADGC-independent
178 sample of 2,724 cognitively normal elderly individuals with at least 2 years of
179 longitudinal clinical follow-up derived from the NIA funded ADCs (data collection
180 coordinated by the National Alzheimer's Coordinating Center). [9] To assess the
181 relationship between polygenic risk and neuropathology, we assessed 2,960 participants
182 from the NIA ADC samples with genotype and neuropathological evaluations. For the
183 neuropathological variables, we examined the Braak stage for neurofibrillary tangles
184 (NFTs) (0: none; I-II: entorhinal; III-IV: limbic, and V-VI: isocortical) [10] and the
185 Consortium to Establish a Registry for Alzheimer's Disease (CERAD) score for neuritic
186 plaques (none/sparse, moderate, or frequent). [11]
187 ADNI: To assess the relationship between polygenic risk and *in vivo* biomarkers, we
188 evaluated an ADGC-independent sample of 692 older controls, mild cognitive
189 impairment and AD participants from the ADNI (see Supplemental Methods). On a
190 subset of ADNI1 participants with available genotype data, we evaluated baseline CSF
191 levels of $A\beta_{1-42}$ and total tau, as well as longitudinal clinical dementia rating-sum of box
192 (CDR-SB) scores. In ADNI1 participants with available genotype and quality-assured
193 baseline and follow-up MRI scans, we also assessed longitudinal sub-regional change in

194 medial temporal lobe volume (atrophy) on 2471 serial T_1-weighted MRI scans (for
195 additional details see Supplemental Methods).
196
197 *Statistical Analysis*
198 We followed three steps to derive the polygenic hazard scores (PHS) for predicting AD
199 age of onset: 1) we defined the set of associated SNPs, 2) we estimated hazard ratios for
200 polygenic profiles, and 3) we calculated individualized absolute hazards (see
201 Supplemental Information for detailed description of these steps).
202 Using the IGAP Stage 1 summary statistics, we first identified a list of SNPs
203 associated with increased risk for AD using significance threshold of $p < 10^{-5}$. Next, we
204 evaluated all IGAP-detected, AD-associated SNPs within the ADGC Phase 1 case-
205 control dataset. Using a stepwise procedure in survival analysis, we delineated the final
206 list of SNPs for constructing the polygenic hazard score. [12-13] In the Cox proportional
207 hazard models, we identified the top AD-associated SNPs within the ADGC Phase 1
208 cohort (excluding NIA ADC and ADNI samples), while controlling for the effects of
209 gender, *APOE* variants, and top five genetic principal components (to control for the
210 effects of population stratification). We utilized age of AD onset and age of last clinical
211 visit to estimate 'age appropriate' hazards [14] and derived a PHS for each participant. In
212 each step of the stepwise procedure, the algorithm selected one SNP from the pool that
213 most improved model prediction (i.e. minimizing the Martingale residuals); additional
214 SNP inclusion that did not further minimize the residuals resulted in halting of the
215 selection process. To prevent over-fitting in the training step, we used 1000x
216 bootstrapping for model averaging and estimating the hazard ratios for each selected

SNPs. We assessed the proportional hazard assumption in the final model using graphical comparisons.

To assess replication, we first examined whether the ADGC Phase 1 derived predicted PHSs could stratify individuals into different risk strata within the ADGC Phase 2 cohort. We next evaluated the relationship between predicted age of AD onset and the empirical/actual age of AD onset using cases from ADGC Phase 2. We binned risk strata into percentile bins and calculated the mean of actual age in that percentile as the empirical age of AD onset.

Because case-control samples cannot provide the proper baseline hazard, [16] we used the previously reported annualized incidence rates by age, estimated from the general United States of America (US) population. [17] For each participant, by combining the overall population-derived incidence rates [17] and genotype-derived PHS, we calculated an individual's instantaneous risk for developing AD, based on their genotype and age (for additional details see Supplemental Information). To independently validate the predicted instantaneous risk, we evaluated longitudinal follow-up data from 2,724 cognitively normal older individuals from the NIA ADC with at least 2 years of clinical follow-up. We assessed the number of cognitively normal individuals progressing to AD as a function of the predicted PHS risk strata and examined whether the predicted PHS-derived incidence rate reflects the empirical/actual progression rate using a Cochran-Armitage trend test.

To assess validity, we examined the association between our PHS and established *in vivo* and pathologic markers of AD neurodegeneration. Using linear models, we assessed whether the PHS correlated with Braak stage for NFTs and CERAD score for

neuritic plaques as well as CSF $A\beta_{1-42}$, and CSF total tau. Using linear mixed effects models, we also investigated whether the PHS was associated with longitudinal CDR-SB score and volume loss within the entorhinal cortex and hippocampus. In all analyses, we co-varied for the effects of age and sex.

RESULTS

PHS: model development, relationship to APOE and independent replication

From the IGAP cohort, we found 1854 SNPs associated with increased risk for AD at a $p < 10^{-5}$. Of these, using the Cox stepwise regression framework, we identified 31 SNPs, in addition to two *APOE* variants, within the ADGC cohort for inclusion into the polygenic model (Table 2). Figure 1 illustrates the relative risk for developing AD using the ADGC case/control Phase 1 cohort. The graphical comparisons among Kaplan-Meier estimations and Cox proportional hazard models indicate the proportional hazard assumption holds for the final model (Figure 1).

To quantify the additional prediction provided by polygenic information beyond *APOE*, we evaluated how PHS modulates age of AD onset in *APOE* ε3/3 individuals. Among these individuals, we found that age of AD onset can vary by more than 10 years, depending on polygenic risk. For example, for an *APOE* ε3/3 individual in the 10^{th} decile (top 10%) of PHS, at a survival proportion of 50%, the expected age for developing AD is approximately 84 years (Figure 2); however, for an *APOE* ε3/3 individual in the 1^{st} decile (bottom 10%) of PHS, the expected age of developing AD is approximately 95 years (Figure 2). Similarly, we also evaluated the relationship between PHS and the

different *APOE* alleles (ε 2/3/4) (Supplemental Figure 1). These findings show that beyond *APOE,* the polygenic architecture plays an integral role in affecting AD risk.

To assess independent replication, we applied the ADGC Phase 1-trained model on independent replication samples from ADGC Phase 2. Using the empirical distributions, we found that the PHS successfully stratified individuals from independent cohorts into different risk strata (Figure 3a). Among AD cases in the ADGC Phase 2 cohort, we found that the predicted age of onset was strongly associated with the empirical (actual) age of onset (binned in percentiles, $r = 0.90$, $p = 1.1 \times 10^{-26}$, Figure 3b).

Predicting population risk of AD onset

To evaluate risk for developing AD, combining the estimated hazard ratios from the ADGC cohort, allele frequencies for each of the AD-associated SNPs from the 1000 Genomes Project and the disease incidence in the general US population, [17] we generated the population baseline-corrected survival curves given an individual's genetic profile and age (Supplemental Figures 2A and 2B). We found that the risk for developing AD as well as the distribution of age of onset is modified by PHS status (Supplemental Figures 2A,B).

Given an individual's genetic profile and age, the corrected survival proportion can be translated directly into incidence rates (Figure 4, Table 3 and Supplemental Table 1). As previously reported in a meta-analysis summarizing four studies from the US general population, [17] the annualized incidence rate represents the proportion (in percent) of individuals in a given risk stratum and age, who have not yet developed AD but will develop AD in the following year; thus the annualized incidence rate represents the

285 instantaneous risk for developing AD conditional on having survived up to that point in
286 time. For example, for a cognitively normal 65 year-old individual in the 80th percentile
287 PHS, the incidence rate would be: 0.29 at age 65, 1.22 at age 75, 5.03 at age 85, and
288 20.82 at age 95 (Figure 4 and Table 3); in contrast, for a cognitively normal 65 year old
289 in the 20th percentile PHS, the incidence rate (per 100 person-years) would be 0.10 at age
290 65, 0.43 at age 75, 1.80 at age 85, and 7.43 at age 95 (Figure 4 and Table 3). As
291 independent validation, we examined whether the PHS predicted incidence rate reflects
292 the empirical progression rate (from normal control to clinical AD) (Figure 5). We found
293 that the PHS predicted incidence was strongly associated with empirical progression rates
294 (Cochrane Armitage trend test, $p = 1.54 \times 10^{-10}$).
295
296 *Association with known markers of AD pathology*
297 We found that the PHS was significantly associated with Braak stage of NFTs (β-
298 coefficient = 0.115, standard error (SE) = 0.024, p-value = 3.9×10^{-6}) and CERAD score
299 for neuritic plaques (β-coefficient = 0.105, SE = 0.023, p-value = 6.8×10^{-6}). We
300 additionally found that the PHS was associated with worsening CDR-Sum of Box score
301 over time (β-coefficient = 2.49, SE = 0.38, p-value = 1.1×10^{-10}), decreased CSF $A\beta_{1-42}$
302 (reflecting increased intracranial Aβ plaque load) (β-coefficient = -0.07, SE = 0.01, p-
303 value = 1.28×10^{-7}), increased CSF total tau (β-coefficient = 0.03, SE = 0.01, p-value =
304 0.05), and increased volume loss within the entorhinal cortex (β-coefficient = -0.022, SE
305 = 0.005, p-value = 6.30×10^{-6}) and hippocampus (β-coefficient = -0.021, SE = 0.0054, p-
306 value = 7.86×10^{-5}).
307

DISCUSSION

In this study, by integrating AD-associated SNPs from recent GWAS and disease incidence estimates from the US population into a genetic epidemiology framework, we have developed a clinically usable, polygenic hazard score for quantifying individual differences in risk for developing AD, as a function of genotype and age. The PHS systematically modified age of AD onset, and was associated with known *in vivo* and pathologic markers of AD neurodegeneration. In independent cohorts, the PHS successfully predicted empirical (actual) age of onset and longitudinal progression from normal aging to AD. Even among individuals who do not carry the ε4 allele of *APOE* (the majority of the US population), we found that polygenic information is useful for predicting age of AD onset.

Using a case/control design, prior work has combined GWAS-associated polymorphisms and disease prediction models to predict risk for AD. [18-19] Rather than representing a continuous process where non-demented individuals progress to AD over time, the case/control approach implicitly assumes that normal controls do not develop dementia and treats the disease process as a dichotomous variable where the goal is maximal discrimination between diseased 'cases' and healthy 'controls'. Given the striking age-dependence of AD, this approach is clinically suboptimal for predicting risk of AD. Building on prior genetic estimates from the general population, [2, 20] we employed a survival analysis framework to integrate AD-associated common variants with established population-based incidence [17] to derive a continuous measure, polygenic hazard score (PHS). From a personalized medicine perspective, for a single non-

330 demented individual, the PHS can estimate individual differences in AD risk across a
331 lifetime and can quantify the yearly incidence rate for developing AD.

332 These findings indicate that the lifetime risk of age of AD onset varies by
333 polygenic profile. For example, the annualized incidence rates (risk for developing AD in
334 a given year) are considerably lower for an 80-year old individual in the 20th percentile
335 PHS relative to an 80-year old in the 99th percentile PHS (Figure 4 and Table 3). Across
336 the lifespan (Supplemental Figure 2B), our results indicate that even individuals with low
337 genetic risk (low PHS) develop AD, but at a later peak age of onset. This suggests that all
338 individuals, irrespective of genotype, would eventually succumb to dementia if they did
339 not die from other causes. Certain loci (including *APOE* ε2) may 'protect' against AD by
340 delaying, rather than preventing, disease onset.

341 Our polygenic results provide important predictive information beyond *APOE*.
342 Among *APOE* ε3/3 individuals, who constitute 70-75% of all individuals diagnosed with
343 late-onset AD, age of onset varies by more than 10 years, depending on polygenic risk
344 profile (Figure 2). At 60% AD risk *APOE* ε3/3 individuals in the 1st decile of PHS have
345 an expected age of onset of 85 whereas for individuals in the 10th decile of PHS, the
346 expected age of onset is greater than 95. These findings are directly relevant to the
347 general population where *APOE* ε4 only accounts for a fraction of AD risk [3] and are
348 consistent with prior work [21] indicating that AD is a polygenic disease where non-*APOE*
349 genetic variants contribute significantly to disease etiology.

350 Using the ADGC phase 2 dataset, we found that the PHS strongly predicted actual
351 age of AD onset in an independent sample indicating the feasibility of using PHS for
352 diagnosing clinical AD. Within the NIA ADC sample, the PHS robustly predicted

longitudinal progression from normal aging to AD illustrating the clinical value of using polygenic information to identify cognitively normal older individuals at highest risk for developing AD (preclinical AD). We found a strong relationship between PHS and increased tau associated NFTs and amyloid plaques suggesting that our genetic marker of disease risk reflects underlying Alzheimer's pathology. The PHS also demonstrated robust associations with CSF $A\beta_{1-42}$ levels, longitudinal MRI measures of medial temporal lobe volume loss and baseline CDR-SB score illustrating that increased genetic risk predicts clinical status and neurodegeneration *in vivo*.

From a clinical perspective, our genetic risk score, based on standard SNP chip arrays, can be used clinically for disease diagnosis, accurate identification of older individuals at greatest risk for developing AD and potentially, for informing management decisions. By providing an accurate, probabilistic assessment as to whether Alzheimer's neurodegeneration is likely to occur, determining a 'genomic profile' of AD may help initiate a dialogue on future planning. Importantly, a continuous, polygenic measure of AD genetic risk may provide an enrichment strategy for prevention and therapeutic trials and could also be useful for predicting which individuals may respond to therapy. Finally, a similar genetic epidemiology framework may be useful for quantifying the risk associated with numerous other common diseases.

There are several limitations to our study. We primarily focused on Caucasian individuals of European descent. Given that AD incidence [20] and genetic risk [22,23] in African-Americans and Latinos is different than in Caucasians, additional work will be needed to develop a polygenic risk model in non-Caucasian populations. The previously reported population annualized incidence rates were not separately provided for males

and females. [17] Therefore, we could not report PHS annualized incidence rates stratified by sex. Finally, we focused on *APOE* and GWAS-detected polymorphisms for disease prediction. Given the flexibility of our genetic epidemiology framework, it can be used to investigate whether a combination of common and rare genetic variants along with clinical, cognitive and imaging biomarkers may prove useful for refining the prediction of AD age of onset.

In conclusion, we have developed, replicated and validated a clinically useful new polygenic hazard score for quantifying the age-associated risk for developing AD. By integrating population based incidence proportion and genome-wide data into a genetic epidemiology framework, we were able to derive hazard estimates whereby an individual could calculate his/her 'personalized' age-specific AD risk, given genetic information. Measures of polygenic risk may prove useful for early detection, determining prognosis, and as an enrichment strategy in clinical trials.

ACKNOWLEDGEMENTS

Drs. Rahul Desikan and Anders Dale had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs. Rahul Desikan (UCSF), Chun Chieh Fan (UCSD), Yunpeng Wang (UCSD and University of Oslo) and Anders Dale (UCSD) conducted and are responsible for the data analysis in this manuscript. The sources of financial and material support had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript. We thank the Shiley-Marcos Alzheimer's Disease Research Center at UCSD and the Memory and Aging Center at UCSF for continued support and the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. This work was supported by grants from the National Institutes of Health (NIH-AG046374, K01AG049152, R01MH100351), the Research Council of Norway (#213837, #225989, #223273, #237250/EU JPND), the South East Norway Health Authority (2013-123), Norwegian Health Association and the KG Jebsen Foundation. Please see Supplemental Acknowledgements for IGAP, NIAGADS, ADGC, ADNI and NACC funding sources.

REFERENCES

1. Kelley AS, McGarry K, Gorges R, MA, Skinner JS. The Burden of Health Care Costs for Patients With Dementia in the Last 5 Years of Life. Ann Intern Med. 2015;163:729-736.
2. Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, Pericak-Vance MA, Risch N, van Duijn CM. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. JAMA. 1997 Oct 22-29;278(16):1349-56.
3. Karch CM, Cruchaga C, Goate AM. Alzheimer's disease genetics: from the bench to the clinic. Neuron 2014;83:11-26.
4. Lambert JC, Ibrahim-Verbaas CA, Harold D et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 2013;45:1452-8.
5. Desikan RS, Schork AJ, Wang Y, et al. Polygenic Overlap Between C-Reactive Protein, Plasma Lipids, and Alzheimer Disease. Circulation. 2015;131:2061-9.
6. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 2007;17:1520-8.
7. Naj AC, Jun G, Beecham GW, Wang LS, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet 2011;43:436-41.
8. McKhann G, Drachman D, Folstein M, et al. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 1984;34:939-44.
9. Beekly DL, Ramos EM, Lee WW, et al. The National Alzheimer's Coordinating Center (NACC) database: the Uniform Data Set. Alzheimer Dis Assoc Disord 2007;21:249-58.
10. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol 1991;82:239-59.
11. Mirra SS, Heyman A, McKeel D, Sumi SM, Crain BJ, Brownlee LM, Vogel FS, Hughes JP, van Belle G, Berg L. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. Neurology 1991;41:479-86.
12. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 2012;44:369-U170.
13. Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. Plos Genet 2013;9.
14. Klein JP, Houwelingen HC, Ibrahim JG, Scheike TH. Handbook of Survival Analysis 2014.
15. Heagerty PJ, Zheng YY. Survival model predictive accuracy and ROC curves. Biometrics 2005;61:92-105.

16. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed: Lippincott Williams & Wilkins; 2008.
17. Brookmeyer R, Gray S, Kawas C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. Am J Public Health 1998;88:1337-42.
18. Escott-Price V, Sims R, Bannister C, et al. Common polygenic variation enhances risk prediction for Alzheimer's disease. Brain. 2015 Oct 21.
19. Yokoyama JS, Bonham LW, Sears RL, et al. Decision tree analysis of genetic risk for clinically heterogeneous Alzheimer's disease. BMC Neurol 2015;15:47.
20. Tang MX, Stern Y, Marder K, et al. The APOE-epsilon4 allele and the risk of Alzheimer disease among African Americans, whites, and Hispanics. JAMA 1998;279:751-5.
21. Sims R, Williams J. Defining the Genetic Architecture of Alzheimer's Disease: Where Next. Neurodegener Dis. 2016;16(1-2):6-11.
22. Tang MX, Cross P, Andrews H, Jacobs DM, et al. Incidence of AD in African-Americans, Caribbean Hispanics, and Caucasians in northern Manhattan. Neurology 2001;56:49-56.
23. Reitz C, Jun G, Naj A, et al. Variants in the ATP-binding cassette transporter (ABCA7), apolipoprotein E 4,and the risk of late-onset Alzheimer disease in African Americans. JAMA 2013;309:1483-92.
24. Jun G, Ibrahim-Verbaas CA, Vronskaya M, et al. A novel Alzheimer disease locus located near the gene encoding tau protein. Mol Psychiatry. 2016 Jan;21:108-17.

Table 1. Demographic data for AD patients and older controls.

	IGAP AD patients	IGAP older controls	ADGC Phase 1 AD patients	ADGC Phase 1 older controls	ADGC Phase 2 AD patients	ADGC Phase 2 older controls
Total N	17,008	37,154	6,409	9,386	6,984	10,972
Mean age (SD) of onset (cases) or assessment (controls)	74.7 (8.0)	68.6 (8.5)	74.7 (7.7)	76.4 (8.1)	73.6 (7.3)	75.7 (8.6)
% Female	63	57	61	59	57.6	60.7
***% APOE* ε4 carriers**	59.0	25.4	51.6	26.7	56.0	28.4

Table 2. Selected 31 SNPs, their closest genes, hazard ratio estimations, and their conditional p values in the final joint model, after controlling for effects of gender and APOE variants.

	Chr	Position	Gene	β	Conditional p in $-\log_{10}$
ε2 allele	19		*APOE*	-0.47	> 15
ε4 allele	19		*APOE*	1.03	> 20
rs4266886	1	207685786	*CR1*	-0.09	2.7
rs61822977	1	207796065	*CR1*	-0.08	2.8
rs6733839	2	127892810	*BIN1*	-0.15	10.5
rs10202748	2	234003117	*INPP5D*	-0.06	2.1
rs115124923	6	32510482	*HLA-DRB5*	0.17	7.4
rs115675626	6	32669833	*HLA-DQB1*	-0.11	3.2
rs1109581	6	47678182	*GPR115*	-0.07	2.6
rs17265593	7	37619922	*BC043356*	-0.23	3.6
rs2597283	7	37690507	*BC043356*	0.28	4.7
rs1476679	7	100004446	*ZCWPW1*	0.11	4.9
rs78571833	7	143122924	*AL833583*	0.14	3.8
rs12679874	8	27230819	*PTK2B*	-0.09	4.2
rs2741342	8	27330096	*CHRNA2*	0.09	2.9
rs7831810	8	27430506	*CLU*	0.09	3.0
rs1532277	8	27466181	*CLU*	0.21	8.3
rs9331888	8	27468862	*CLU*	0.16	5.1
rs7920721	10	11720308	*CR595071*	-0.07	2.9
rs3740688	11	47380340	*SPI1*	0.07	2.8
rs7116190	11	59964992	*MS4A6A*	0.08	3.9
rs526904	11	85811364	*PICALM*	-0.20	2.3
rs543293	11	85820077	*PICALM*	0.30	4.2
rs11218343	11	121435587	*SORL1*	0.18	2.8
rs6572869	14	53353454	*FERMT2*	-0.11	3.0
rs12590273	14	92934120	*SLC24A4*	0.10	3.5
rs7145100	14	107160690	*abParts*	0.08	2.0
rs74615166	15	64725490	*TRIP4*	-0.23	3.1
rs2526378	17	56404349	*BZRAP1*	0.09	4.9
rs117481827	19	1021627	*C19orf6*	-0.09	2.5
rs7408475	19	1050130	*ABCA7*	0.18	4.3
rs3752246	19	1056492	*ABCA7*	-0.25	8.4
rs7274581	20	55018260	*CASS4*	0.10	2.1

Table 3. Predicted annualized incidence rate (per 100 person-years) by age using polygenic hazard scores.

Age	Population Baseline*	PHS 1 percentile (95% CI)	PHS 20th percentile (95% CI)	PHS 80th percentile (95% CI)	PHS 99th percentile (95% CI)	*APOE* ε4+ (95% CI)	*APOE* ε4- (95% CI)
60	0.08	0.02 (0.01,0.03)	0.04 (0.01,0.08)	0.15 (0.04, 0.27)	0.61 (0.16, 1.06)	0.19 (0.18, 0.20)	0.06 (0.06, 0.7)
65	0.17	0.04 (0.01,0.06)	0.09 (0.03, 0.16)	0.32 (0.09, 0.54)	1.24 (0.33,2.15)	0.38 (0.36, 0.40)	0.13 (0.12, 0.13)
70	0.35	0.07 (0.02,0.13)	0.19 (0.05,0.32)	0.64 (0.18, 1.10)	2.53 (0.68, 4.38)	0.78 (0.74, 0.82)	0.26 (0.25, 0.27)
75	0.71	0.15 (0.05,0.19)	0.38 (0.11,0.65)	1.30 (0.36,2.25)	5.15 (1.38, 8.91)	1.58 (1.51, 1.66)	0.53 (0.52, 0.55)
80	1.44	0.31 (0.26,0.26)	0.77 (0.22,1.32)	2.65 (0.74, 4.57)	10.47 (2.81, 18.13)	3.22 (3.06, 3.38)	1.08 (1.05, 1.11)
85	2.92	0.63 (0.19,1.07)	1.57 (0.45, 2.68)	5.39 (1.50, 9.29)	21.30 (5.72, 36.88)	6.55 (6.23, 6.87)	2.2 (2.13, 2.27)
90	5.95	1.28 (0.38,2.18)	3.19 (0.91, 5.46)	10.97 (3.05, 18.89)	43.32 (11.63, 75.00)	13.33 (12.68, 13.98)	4.48 (4.34, 4.61)
95	12.1	2.61 (0.78,4.44)	6.48 (1.85, 11.10)	22.31 (6.20, 38.43)	88.11 (23.66, 100.00)	27.11 (25.79, 28.43)	9.1 (8.83, 9.38)

* US community-sampled population incidence proportion (% year) reported by reference 17.

APOE ε4+ refers to individuals with at least one copy of the ε4 allele of *APOE*; *APOE* ε4- refers to individuals with no copies of the ε4 allele of *APOE*

493 **FIGURE LEGENDS**

494

495 **Figure 1.** Kaplan-Meier estimates and Cox proportional model fits from the case-control

496 ADGC phase 1 dataset, excluding NACC and ADNI samples. The proportional hazard

497 assumptions were checked based on the graphical comparisons between Kaplan-Meier

498 estimation and Cox proportional hazard models. 95% confidence intervals of Kaplan-

499 Meier estimation are also demonstrated. The baseline hazard (gray line) in this model is

500 based on the mean of ADGC data.

501

502 **Figure 2.** Kaplan-Meier estimates and Cox proportional model fits among *APOE* □ 3/□

503 3 individuals in ADGC phase 1 dataset, excluding NACC and ADNI samples.

504

505 **Figure 3. (a)** Risk stratification in ADGC phase 2 cohort, using PHS derived from

506 ADGC phase 1 dataset. **(b)** Predicted age of AD onset as a function of empirical age of

507 AD onset among cases in ADGC phase 2 cohort. Prediction is based on the final survival

508 model trained in the ADGC phase 1 dataset.

509

510 **Figure 4.** Annualized incidence rates showing the instantaneous hazard as a function of

511 PHS percentiles and age. The gray line represents the population baseline estimate.

512

513 **Figure 5.** Empirical progression rates observed in the NIA ADC longitudinal cohort as a

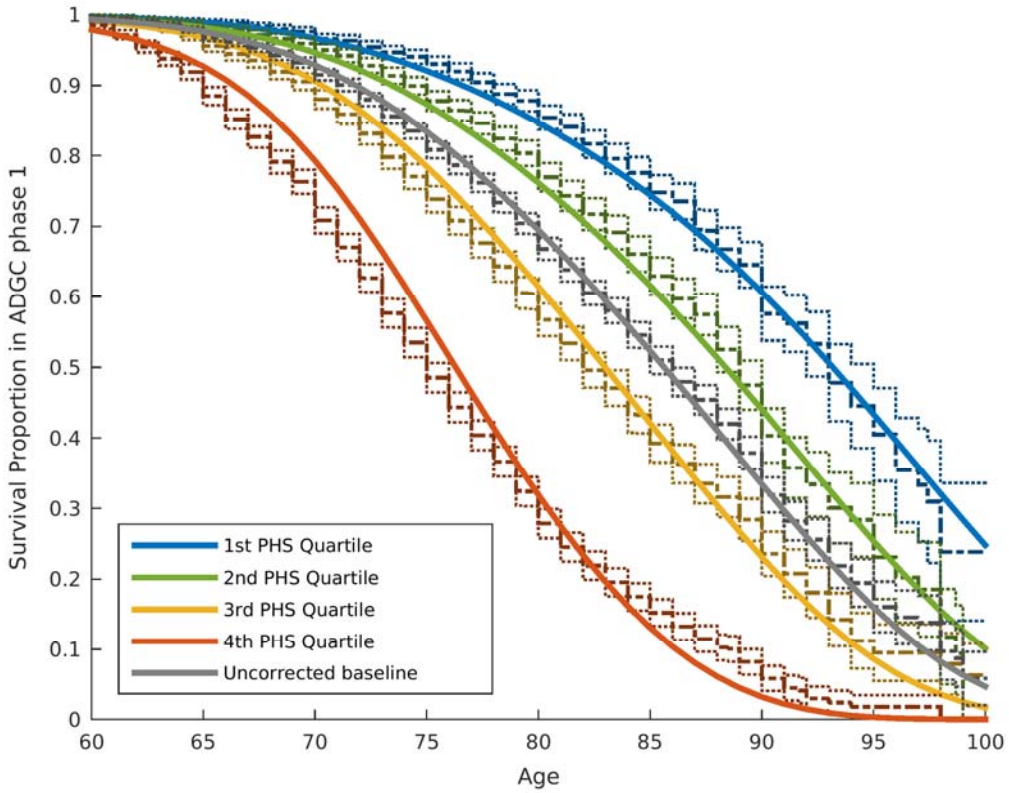
514 function of predicted incidence. CA = Cochrane-Armitage test

515

516

517

518 **Figure 1.**

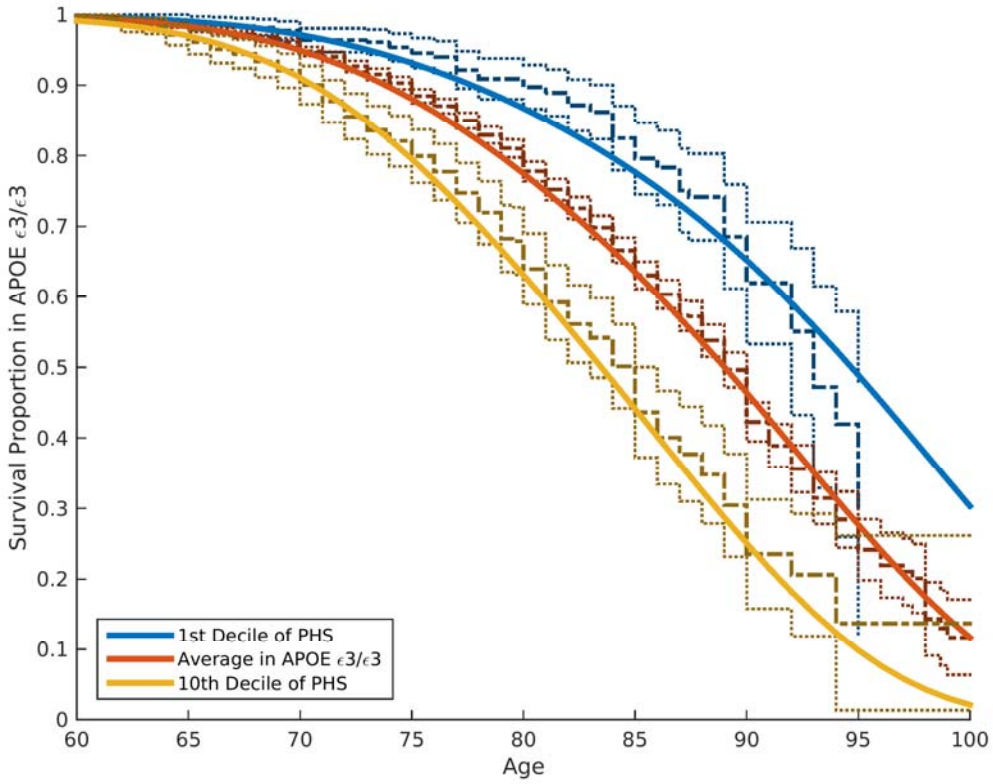


519

520

521

522 **Figure 2.**



523

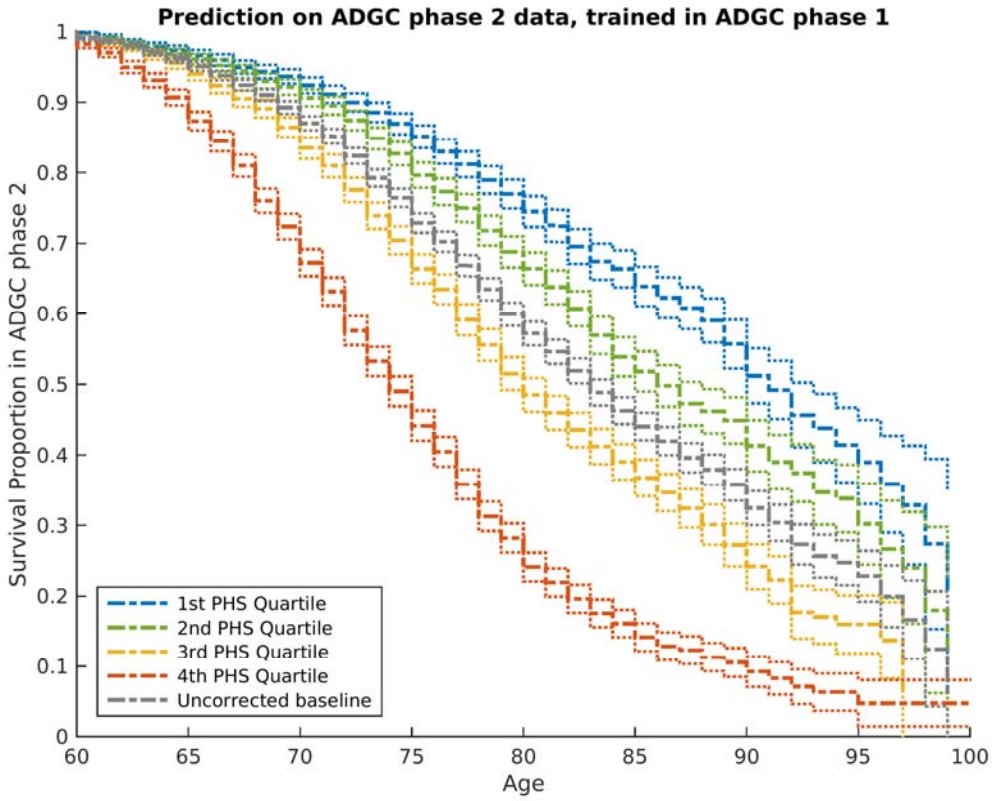
524

525

526

527

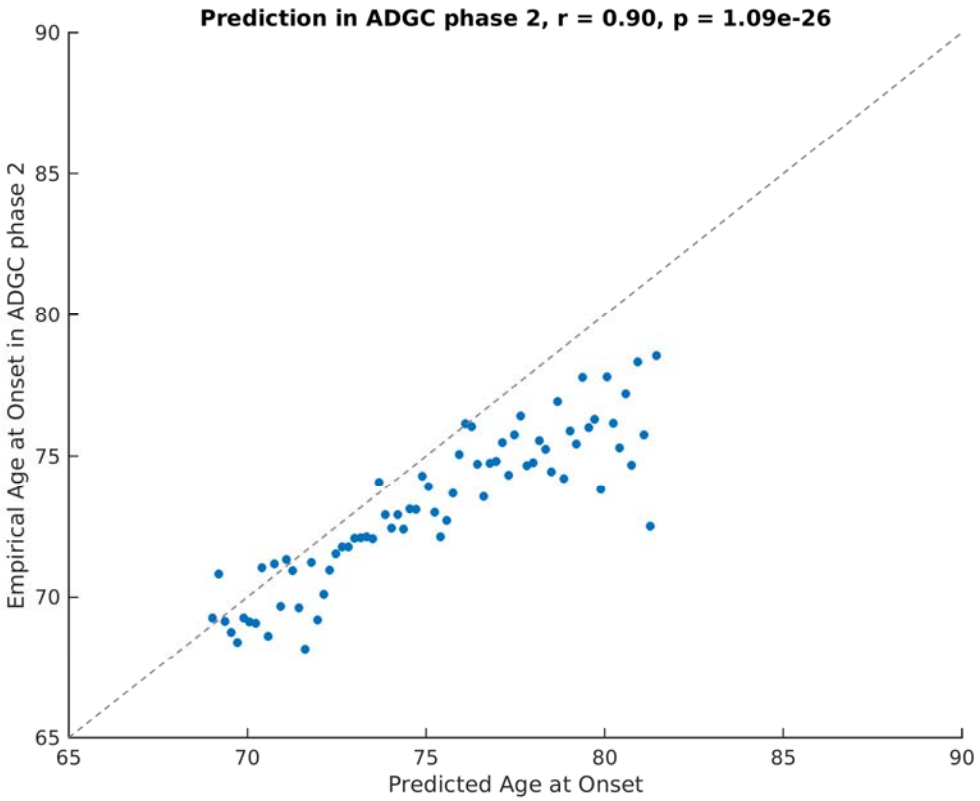
528 **Figure 3a**



529

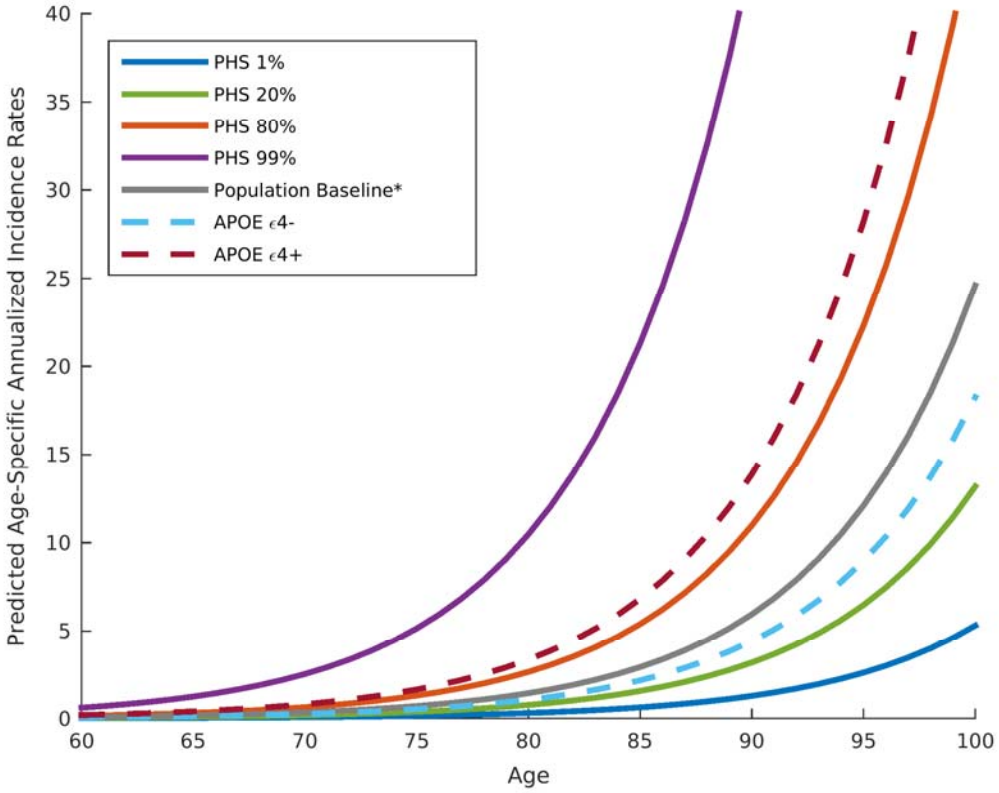
530
531
532

533 **Figure 3b.**



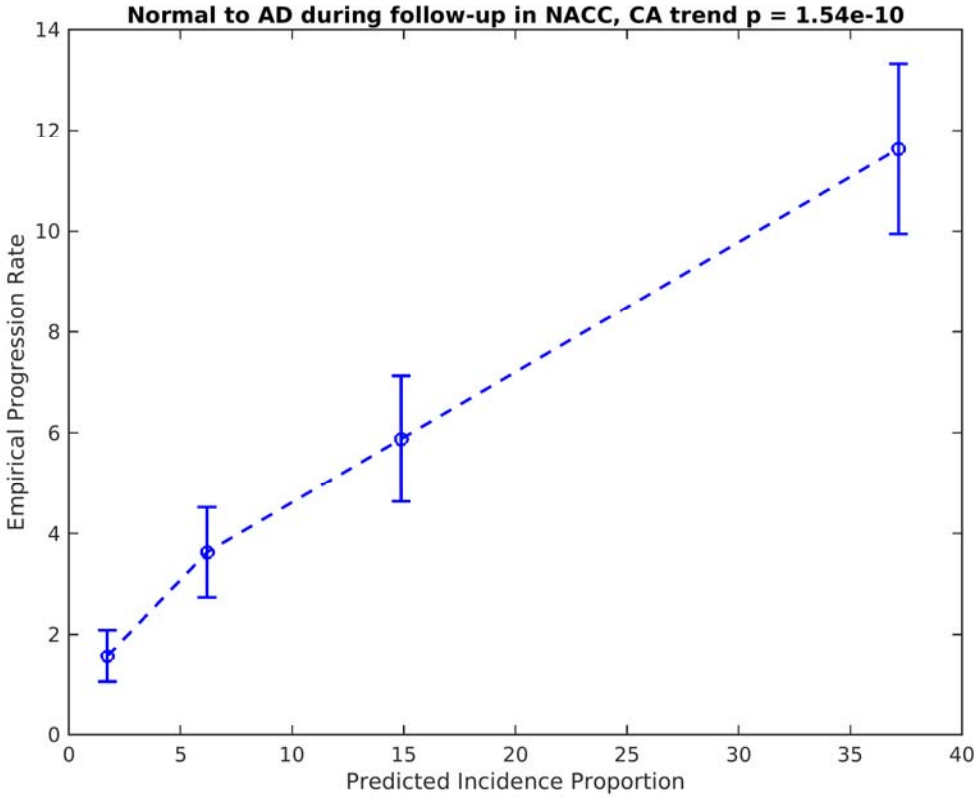
534
535
536
537

538 **Figure 4.**



539
540
541

542 **Figure 5.**



543