Nomograms of Human Hippocampal Volume Shifted by Polygenic Scores

Mohammed Janahi[1,2], Leon Aksman[3], Jonathan M Schott[4], Younes Mokrab[2,5] and Andre Altmann[1] for the Alzheimer's Disease Neuroimaging Initiative[*]

1 Centre for Medical Image Computing (CMIC), Department of Medical Physics and Biomedical Engineering, University College London, United Kingdom

2 Medical and Population Genomics Lab, Human Genetics Department, Research Branch, Sidra Medicine, Doha, Qatar

3 Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States of America

4 Dementia Research Centre (DRC), Queen Square Institute of Neurology, University College London, United Kingdom

5 Department of Genetic Medicine, Weill Cornell Medicine-Qatar, Doha, Qatar

Abstract

Nomograms are important clinical tools applied widely in both developing and aging populations. They are generally constructed as normative models identifying cases as outliers to a distribution of healthy controls. Currently used normative models do not account for genetic heterogeneity. Hippocampal Volume (HV) is a key endophenotype for many brain disorders. Here, we examine the impact of genetic adjustment on HV nomograms and the translational ability to detect dementia patients. Using imaging data from 35,686 healthy subjects aged 44 to 82 from the UK BioBank (UKB), we built HV nomograms using gaussian process regression (GPR), which - compared to a previous method - extended the application age by 20 years, including dementia critical age ranges. Using HV Polygenic Scores (HV-PGS), we built genetically adjusted nomograms from participants stratified into the top and bottom 30% of HV-PGS. This shifted the nomograms in the expected directions by ~100 mm^3 (2.3% of the average HV), which equates to 3 years of normal aging for a person aged ~65. Clinical impact of genetically adjusted nomograms was investigated by comparing 818 subjects from the AD neuroimaging (ADNI) database diagnosed as either cognitively normal (CN), having mild cognitive impairment (MCI) or Alzheimer's disease patients (AD). While no significant change in the survival analysis was found for MCI-to-AD conversion, an average of 68% relative decrease was found in intra-diagnostic-group variance, highlighting the importance of genetic adjustment in untangling phenotypic heterogeneity.

Introduction

Brain imaging genetics is a rapidly evolving area of neuroscience combining imaging, genetic, and clinical data to gain insight into normal and diseased brain morphology and function [1]. Normative modelling is an emerging method in neuroscience, aiming to identify cases as outliers to a distribution of healthy controls and was shown to have potential to improve early diagnosis, progression models, and risk assessment[2-5]. Where conventional case-control studies generally require both cases and controls to be well clustered, normative models work well even when cases are not clustered or overlap with controls. Nomograms are a common implementation of normative models and have been used as growth charts of brain volumes across age in both developing and aging populations[6-8].

Normative modelling identifies cases by their deviation from normality, however, genetics shapes what is 'normal'. Heritability studies have found that whole brain volume is $90\% \pm 4.8\%$ heritable[9], hippocampal volume is $75\% \pm 5\%$[10-12], and other cortical brain areas between 34% and 80%[13,14]. Genome-wide association studies (GWAS) have identified genome wide significant variants that explain $13\% \pm 1.5\%$ of the variation in hippocampal volume (HV) [15], $34\% \pm 3\%$ in total cortical surface area, and $26\% \pm 2\%$ in average cortical thickness[16]. The gap between estimates from GWAS hits and formal heritability estimates (termed the 'missing heritability')[17] implies that less significant variants also have an influence and that it may be captured through polygenic scores (PGS)[18-20]. In this work we demonstrate the impact of accounting for polygenic effects in normative modelling of HV.

Damage to the hippocampus (which is integral to memory processes[21]) has been associated with major depressive disorder[22], schizophrenia[23] , Epilepsy[24], and Alzheimer's disease (AD)[25]. AD is a global health burden: 7% percent of people over 60 are diagnosed with dementia[26] of which AD accounts for 70%[27]. The pathophysiological processes underlying AD, namely amyloid and tau pathology accumulation, are thought to precede brain atrophy, which typically starts in the hippocampus and medial temporal lobe and then spreads throughout the neocortex[27].

The normal variation of HV is of great clinical interest as the early and often prominent hippocampal atrophy seen in AD creates a need for early diagnosis and disease tracking.

107   Many studies have examined HV across age[28,29], for example, a recent study by Nobis et.

108   (2019)[30] produced HV nomograms from UK Biobank (UKB) for use in clinical settings. It is

109   important to note that some of the variation in the normative models can be explained by

110   the clear impact of genetics on HV[15,31]. Thus far, the few attempts at including genetics in

111   the construction of HV nomograms have focussed on disease related variants. For instance,

112   two recent studies examined the impact of the AD-associated *APOE* gene[32,33], showing that

113   APOE4/4 carriers had significantly lower HV trajectories. This effect is likely driven by AD-

114   related disease processes since APOE4/4 carriers have a 10-fold risk of developing AD[34,35].

115   However, the genetic impact on variation in HV in healthy population remains

116   underexamined in the context of nomograms. In this work, we close this gap. We built HV

117   nomograms using a gaussian process regression (GPR) method (Figure 1A). We then

118   computed a PGS of HV for subjects in our cohort and built genetically adjusted nomograms

119   (Figure 1B). We found that genetic adjustment did in fact shift the nomograms and that,

120   because the model requires no smoothing, our GPR nomograms provided an extended age

121   range compared to previous methods.
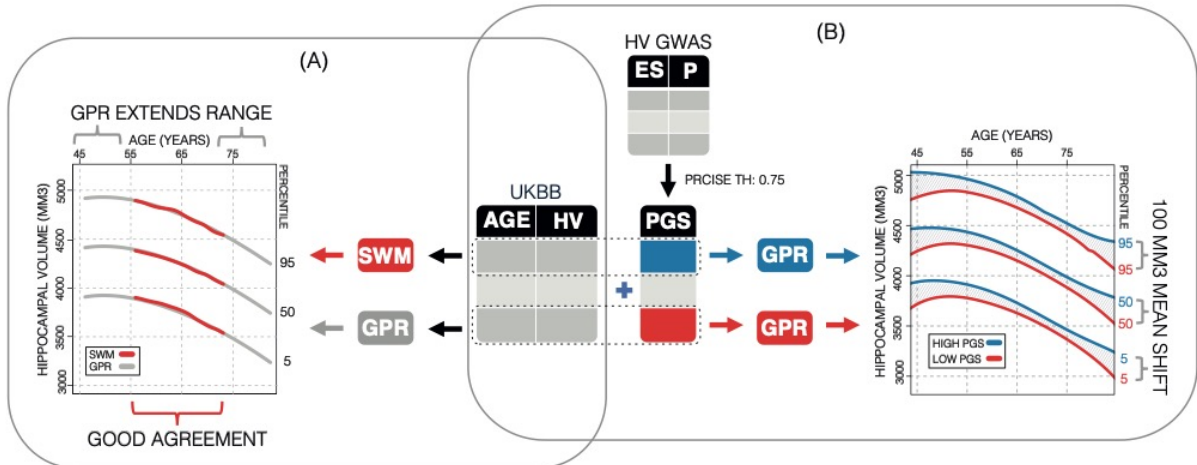
Figure 1: Study Overview. (A) Using 35,686 subjects from the UK BioBank, we generate
nomograms using two methods: a previously reported Sliding Window Method (SWM), and
Gaussian Process Regression (GPR). We find that GPR is more data efficient than the SWM
and can extend the nomogram into dementia critical age ranges. (B) Using a previously
reported genome wide association study, we generate polygenic scores (PGS) for the
subjects in our UK BioBank table. We then stratify the table by PGS and generate
nomograms for the top and bottom 30% of samples separately. We find the genetic
adjustment differentiates the nomograms by an average of 100 mm3, which is equivalent to
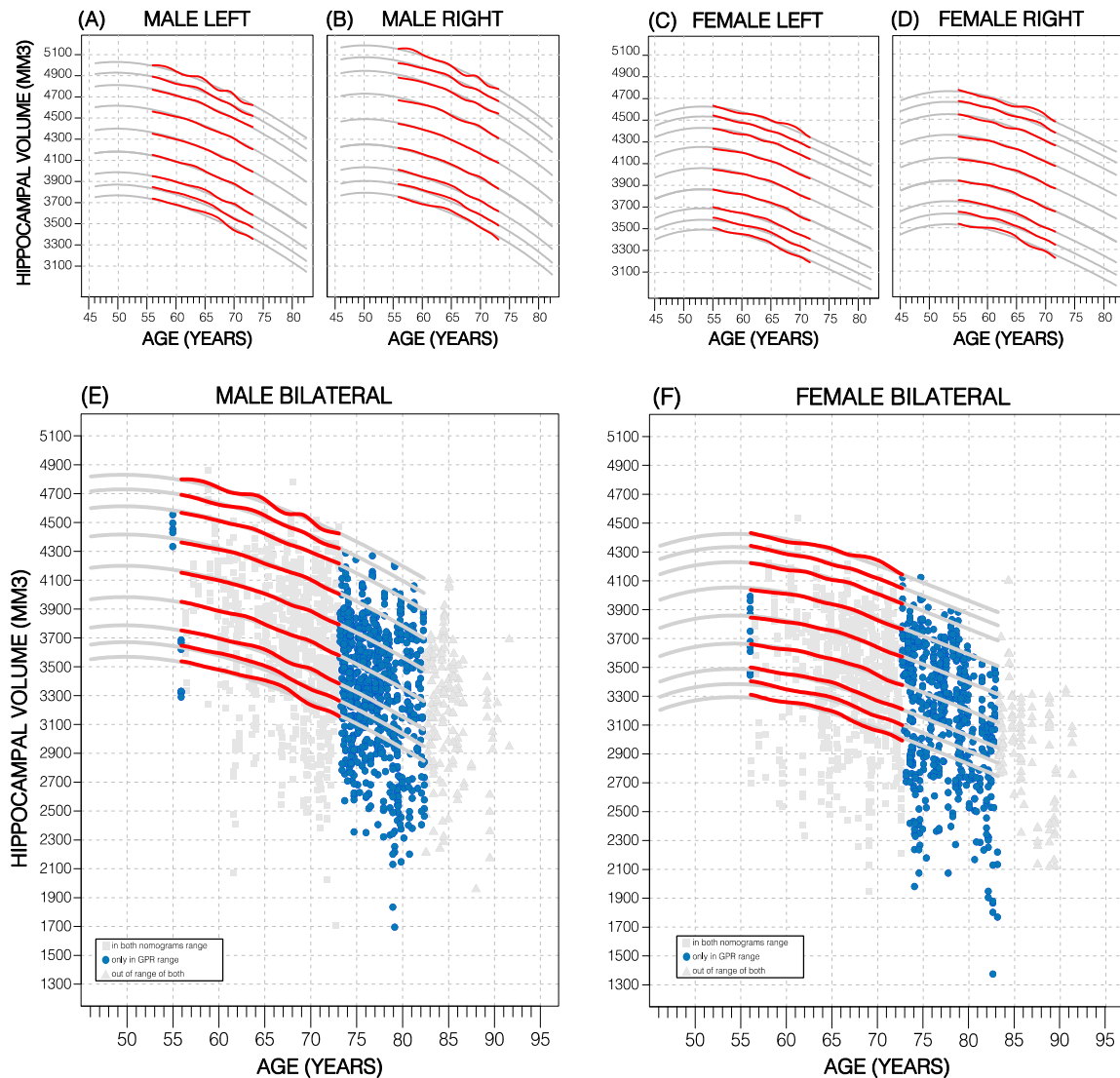about 3 years of normal aging for a 65-year-old.

136   **Results**

137   In the UKB sample 453 subjects were excluded for various conditions, 3497

138   for genetic ancestry, and 28 subjects were outliers: leaving a total of 35,686 subjects. In the

139   ADNI application dataset, 26 subjects were excluded for genetic ancestry, and 314 based on

140   HV quality scores: leaving 818 subjects.

141

142   *SWA vs GPR for nomogram estimation*

143   Nomograms of healthy subjects generated using the SWA and GPR method displayed similar

144   trends (Figure 2; Figure 2 – Figure Supplement 2). However, GPR nomograms spanned the

145   entire training dataset age range (45-82 years) compared to the SWA (52-72 years). This is

146   primarily because the SWA is a non-model-based approach that requires smoothing to avoid

147   edge effects, and a gaussian smoothing window of width 20 was used [30]. This extension

148   allowed 86% of all diagnostic groups from the ADNI to be evaluated versus 56% in the SWA

149   Nomograms (Figure 2; Figure 2 – Figure Supplement 2). Furthermore, our GPR nomograms

150   confirmed previously reported trends: Overall, the average 50th percentile in male

151   nomograms (4162 ± 222) was higher than the female nomograms (3883 ± 170), and

152   within each sex, right HV was larger than left HV (Figure 2; Figure 2 – Figure Supplement 2).

153   We also observed that along the 50th percentile, male HV declined faster (-20.3 mm^3/year)

154   than female HV (-14.6 mm^3/year). Additionally, in GPR nomograms, HV peaks in women at

155   age 53.5 years with a less pronounced peak in males at 50 years (Figure 2; Figure 2 – Figure

156   Supplement 2). Training the GRP model with 16,000 samples took ~1 hour on a consumer

157   grade machine (2.3 GHz 8-Core Intel Core i9).

Figure 2: Comparing Nomogram Generation Methods. Nomograms produced from healthy UKB subjects using the sliding window approach (SWA) (red lines) and gaussian process regression (GPR) method (grey lines) show similar trends. both left hemisphere nomograms (A, C) are lower than their right counterparts (B, D). Male nomograms are higher than female nomograms (A vs C) and (B vs D). Female HV shows a peak at 53.5 years of age, while male HV shows a less prominent peak at 50 years of age. SWA and GPR show good agreement, while GPR enables a 10-year nomogram extension in either direction. The benefits of this extension can be seen with scatter plots of ADNI subjects of all diagnoses overlayed (E, F). The extended age range of the GPR nomograms (45-82 years) enables the evaluation of an additional 43% of male data (E) and 34% of female data (F) (turquoise circles). A similar figure with only the Cognitively Normal ADNI subjects can be found in Figure 2 – Figure Supplement 2

174   *Polygenic Score for Hippocampal Volume*

175   The calculated PGS, based on an earlier GWAS for average bilateral hippocampal volume[15],

176   as expected, showed a strong correlation with HV in the UKB data. Overall, the PGSs showed

177   a significant positive correlation with HV across all p-value thresholds and training sample

178   subsets ($p<2.7E-24$; Table 1). PGSs explained more variance in males versus females.

179   Furthermore, PGSs did not show detectable differences in left versus right HV; and

180   explained the most variance in mean bilateral HV (Table 1, Figure 3 – Data Source 1). In all

181   tested settings, the explained variance (R^2) by the PGS across p-value threshold was similar:

182   with one peak at the 1E-7 threshold (capturing few but very significant SNPs), a higher peak

183   at the 0.75 threshold (capturing many SNPs with mostly small effect sizes) (Figure 3). For the

184   ADNI dataset, this distribution increased with the threshold. When investigating mean HV

185   across percentile of PGS at the 0.75 threshold (highest R^2), the top and bottom 20% of

186   scores accounted for 41% of the variance in HV (Figure 3); with similar values observed

187   across thresholds in both datasets (Figure 3 – Figure Supplement 1, 2).

188
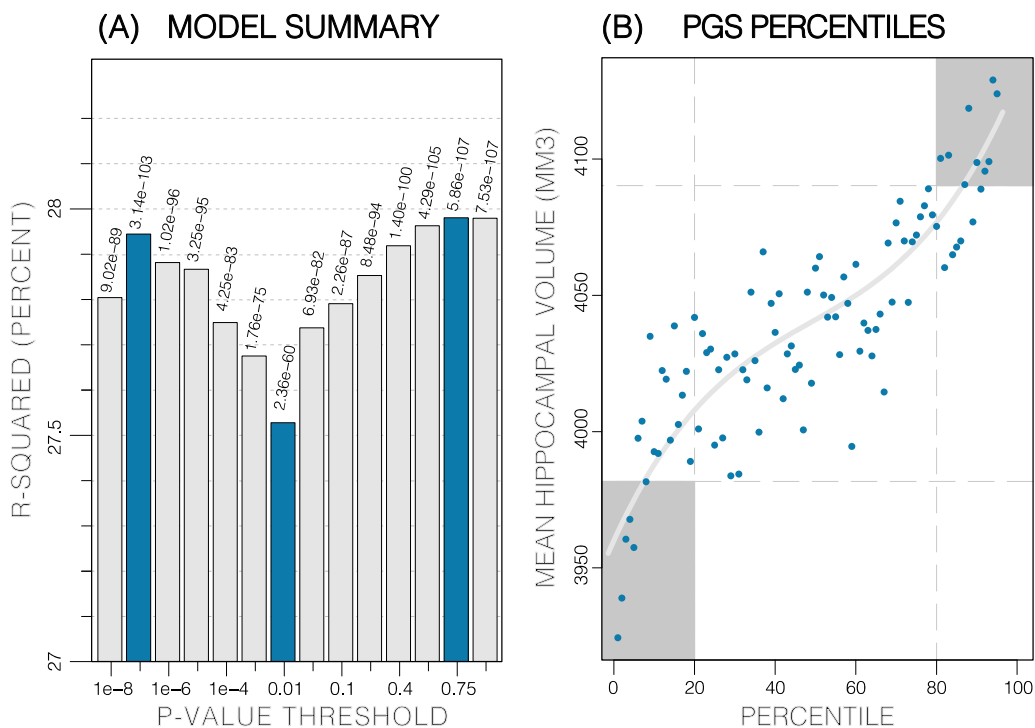
189
190
191   **Figure 3: Summary of PGS models.** Polygenic Risk Score in models of mean HV across both
192   sexes. (a) R^2 of linear models across increasing p-value thresholds. All models are of bilateral
193   HV and account for age, sex, and top 10 genetic principal components. The minimum R^2 on

194   the y-scale is the R^2 of the models without any PGS. (b) distribution of mean HV across

195   percentiles of PGS. Excluding the top and bottom 20% of percentiles reduces the variance by

196   49% (darker grey areas). Fitting a cubic polynomial to the means produces the grey line.

197

198   **Table 1: Association between Polygenic Scores (PGS) and Hippocampal Volume (HV).**

199   Linear models were built for HV (left; right; bilateral) using PGS across cohorts (male;

200   female; both) at three representative p-value thresholds (1E-7; 0.01; 1). p-values of the

201   slope were significant across all categories, with the lowest being associated with the

202   threshold value of 1 in all but a single case (both/right). Variance explained (R^2) increased

203   from left to right to bilateral volumes and increased from female to male to both.

204

Gender	PGS threshold	LEFT			RIGHT			BILATERAL		
		Slope ($\times 10^{-2}$)	p-value	R^2	Slope ($\times 10^{-2}$)	p-value	R^2	Slope ($\times 10^{-2}$)	p-value	R^2
FEMALE	1E-7	10	1.8E-46	13%	9.4	2.4E-45	14%	11	1.4E-51	15%
	0.01	8.2	2.7E-26	13%	7.6	1.0E-27	13%	8.7	3.2E-30	14%
	1	11	**9.4E-54**	13%	9.62	**1.5E-48**	14%	11	**1.6E-57**	15%
MALE	1E-7	8.2	1.4E-35	18%	7.5	2.6E-35	18%	9.2	4.1E-40	20%
	0.01	7.8	3.8E-29	18%	6.8	3.8E-27	18%	8.6	7.8E-32	20%
	1	9.4	**3.2E-48**	18%	8.0	**4.7E-43**	18%	10	**9.1E-52**	20%
BOTH	1E-7	8.4	8.1E-90	25%	7.9	**6.4E-93**	26%	9.3	3.1E-103	28%
	0.01	7.4	9.3E-54	24%	6.7	3.3E-53	26%	8	2.3E-60	28%
	1	9.6	**2.1E-99**	25%	8.3	1.8E-89	26%	10	**7.5E-107**	28%

205

206   Slope = beta coefficient for PGS in the linear mode; p-value for the slope; R^2 = variance

207   explained by the linear model

208

209

210   *Genetics stratified Nomograms*

211   We will focus on the p-value threshold of 0.75 as it achieved best or close to-best

212   performance overall (Figure 3 – Data Source 1). Genetics had a clear effect on the

213   nomograms: the high-PGS nomograms were shifted upwards while the low-PGS nomograms

214   were shifted downwards; an effect which could be observed at both the model and data

215   level (Figure 4; Figure 4 – Figure Supplement 3), both by around 1.2% of the average HV (50

216   mm^3). Thus, the difference between high and low PGS nomograms was ~2.3% of the

217   average HV (100 mm^3). An ANOVA test of the percentiles produced with the adjusted vs

218   unadjusted nomograms revealed that the groups were significantly different to each other

219   (F>19; P<8.04e-6; Table 2). The HV peak previously observed at 50 years in males was less

220   pronounced in the high-PGS nomogram and more so in the low-PGS nomogram (Figure 4,

221   Figure 4 – Figure Supplement 1). Adjusting nomograms using ICV and AD PGSs, instead of

222   HV PGS, did not result in nomograms that were meaningfully different from the non-

223   adjusted nomograms (Figure 4 – Figure Supplement 2).
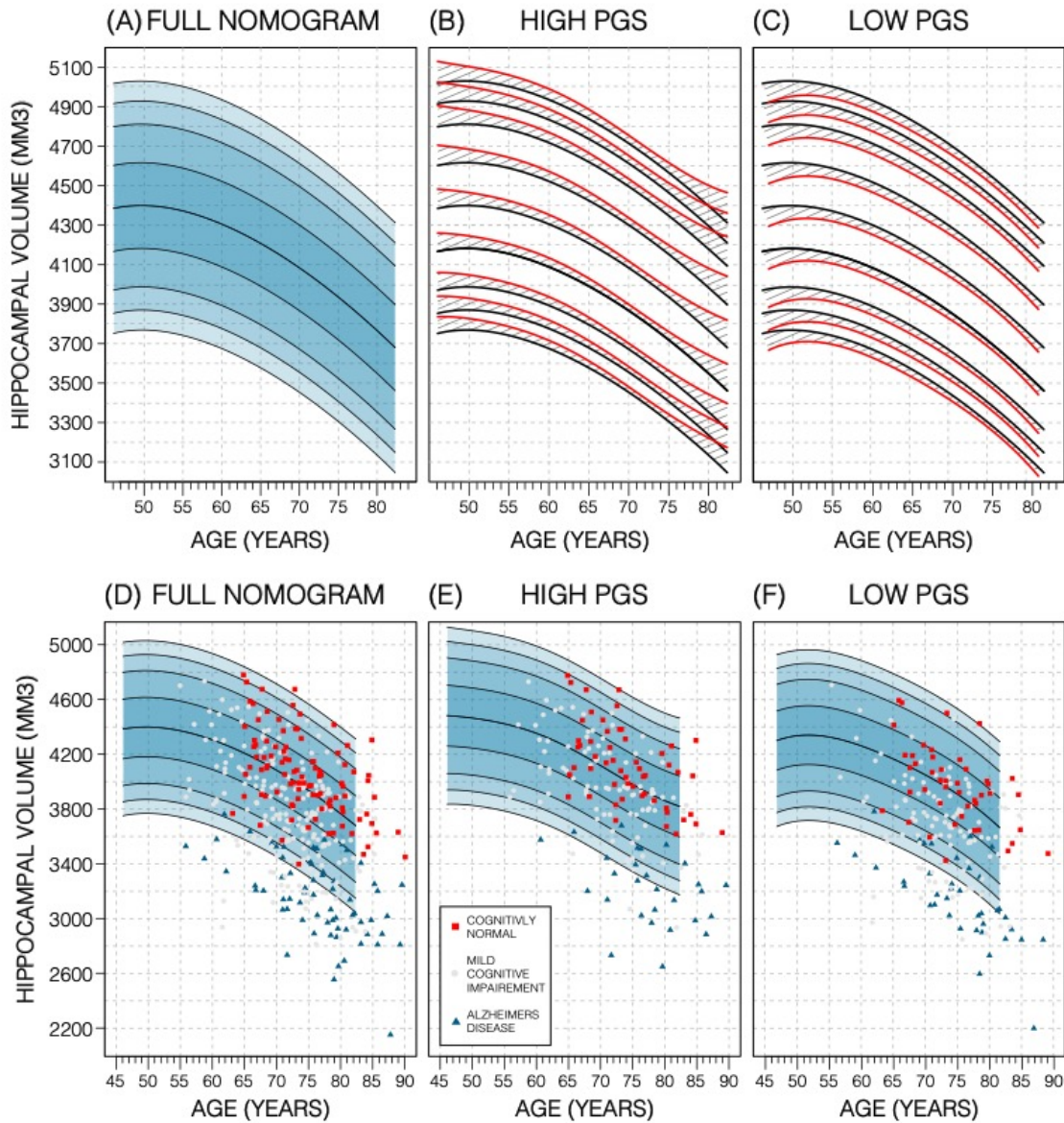
224
225   **Figure 4: Genetically Adjusted Nomograms.** Results of genetic adjustment in bilateral male
226   hippocampal volume (HV). (A, D) Nomograms of bilateral hippocampal volume (HV)
227   generated from all male UKB samples overlayed with male ADNI samples. CN samples (red
228   squares) centre around the 50th percentile, AD samples (turquoise triangles) lie mostly
229   below the 2.5[th] percentile, and MCI samples (grey circles) span both regions. (B, E)
230   Nomograms generated using only high PGS samples (top 30%) was shifted upward (red
231   lines) compared to the original (black lines) by an average of 50 mm^3 (1.2% of mean HV).
232   Plotting the high PGS ADNI samples (top 50%) slightly improves intra-group variance. (C, F)
233   similar results are seen in low PGS samples. Note, the black lines in panels (B, C) are the
234   same as the nomogram in panel (A) and similarly the red lines in panel (B, C) are same as the
235   nomogram in panels (E, F).
236

Table 2: Results of ANOVA tests of UKB HV Percentiles produced with genetically adjusted and unadjusted nomograms.

SEX	STRATA	DF	SUM SQ	F-VALUE	P-VALUE
MEN	HIGH	1	18786	22.84	1.8e-06 ***
MEN	LOW	1	16407	19.96	8.04e-06 ***
WOMEN	HIGH	1	27068	32.92	9.97e-09 ***
WOMEN	LOW	1	30103	36.94	1.28e-09 ***

External Evaluation on ADNI data

In the ADNI dataset we investigated whether the shift in genetically adjusted nomograms could be leveraged for improved diagnosis. Using the non-adjusted nomogram, cognitively normal (CN) participants (n = 225) had a median bilateral HV percentile of 61% (±25% SD), Mild Cognitive Impairment (MCI) participants (n = 391) had 25% (±26% SD), and Alzheimer's Disease (AD) participants (n = 121) had 1% (±9% SD) (Figure 5). Visual inspection revealed that while CN participants were spread across the quantiles, AD participants lay mostly below the 2.5% quantile, and MCI participants spanned the range of both CN and AD participants (Figure 4). Bisecting the samples by PGS showed that high PGS CN samples had median percentiles of 65% (±27% SD) and low PGS had 54% (±26% SD). When comparing the same samples against the genetically adjusted nomograms instead, high PGS CN samples had 60% (±26% SD) and low PGS had 59% (±26% SD). Thus, reducing the gap between high and low PGS CN participants by 9% (from 10% to 1%; a 90% relative reduction). Similar analysis showed a reduction in MCI participants by 10% (60% relative reduction), and 0.5% (56% relative reduction) in AD participants. The above effects persisted across most strata (i.e., sex and hemisphere) (Figure 5; Figure 5 – Data Source 1).
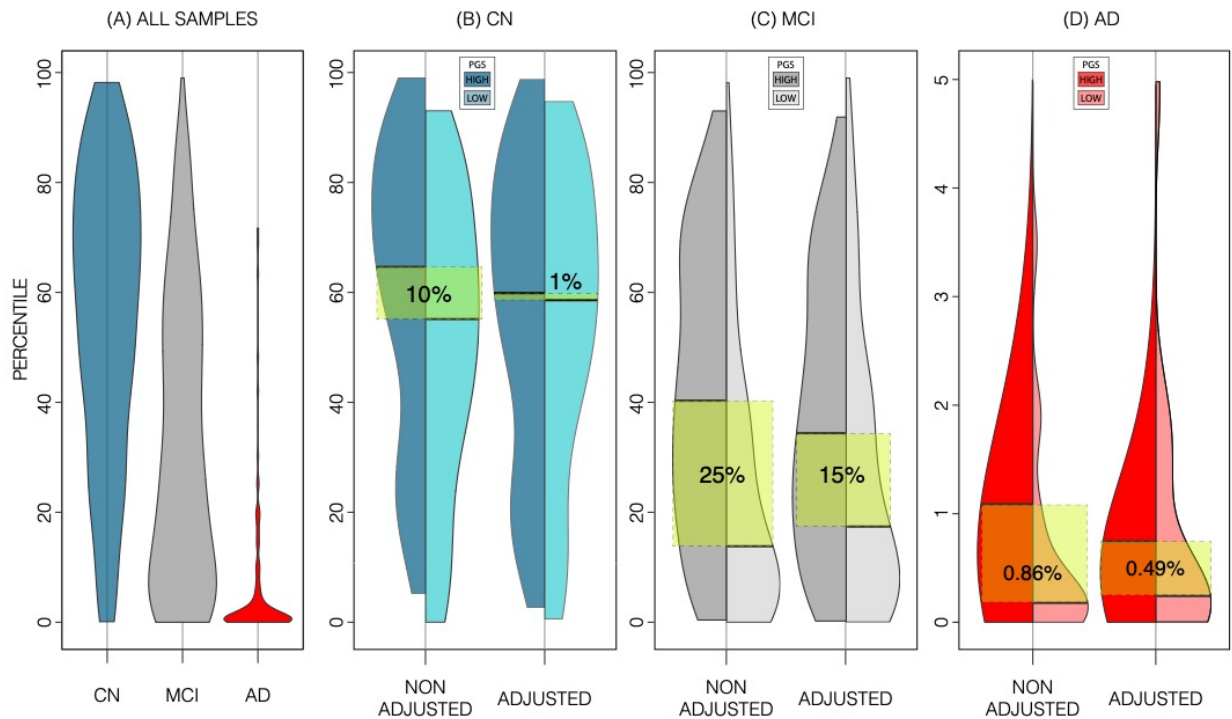
Figure 5: ADNI Dataset Percentiles in Genetically Adjusted/Non-Adjusted Nomograms.
Plotting the percentile distribution of the different diagnostic groups across adjusted and non-adjusted nomograms reveals that genetic adjustment increases group cohesiveness. (A) The percentile distributions of the different diagnostic groups against the non-adjusted nomograms. (B) In CN samples for example, when plotting against the non-adjusted nomogram (left adjoined boxplots), the median percentile of the top 30% of samples (darker turquoise) was 65%, while the median for the lower 30% of samples (lighter turquoise) was 54%. When using the genetically adjusted nomogram instead (right adjoined boxplots), those median percentiles become 60% and 59% respectively; a 90% relative reduction. Similar results can be seen with MCI (C) and AD (D) samples, with 60% and 56% relative reduction respectively.

Longitudinal Evaluation

We also investigated whether genetically adjusted nomograms provided additional accuracy in distinguishing stable (n = 299) from MCI-to-AD progressing subjects (n = 83). With the non-adjusted nomogram, progressing MCI participants had a mean HV percentile of 11% and stable participants had 29% (Figure 6). Using the genetically adjusted nomograms, they had 10% and 28%, respectively. Cox proportional hazards models of percentiles obtained using both nomograms revealed little difference between the two in terms of clinical conversion: both models resulted in a hazard ratio of 0.97 for percentile in nomogram (beta of -0.03 at p-value < 4.87e-08); confirming that participants within lower HV percentiles where more likely to convert earlier.

12

282
283
284
285
286
287
288
289
290
291
292
293
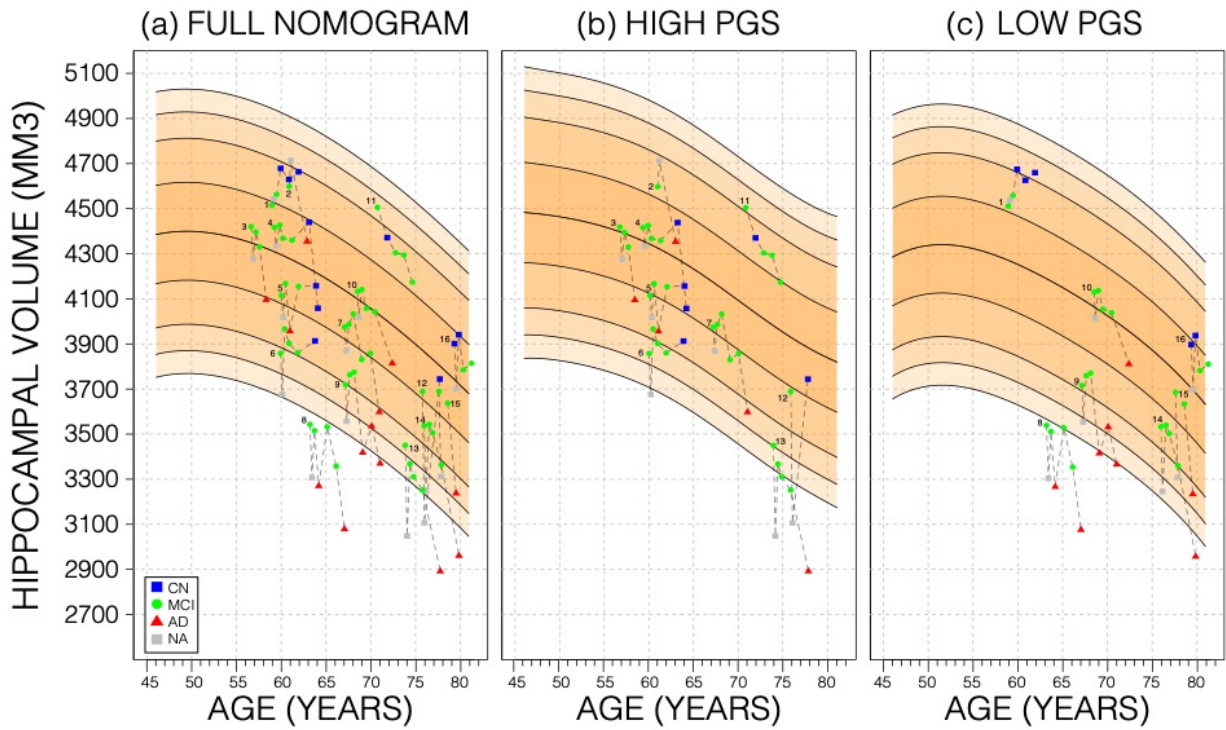294
295
296
297
298
299
300
301
302
303
304
305

Figure 6: Longitudinal Analysis. A selection of MCI samples longitudinal data plotted against nomograms of male mean HV. (a) all selected samples plotted against a non-adjusted nomogram. Lines connect visits of the same sample with diagnosis at each visit shown: CN as blue squares; MCI as green dots, AD as red triangles, and no diagnosis (NA) as grey squares. (b) samples from (a) with high PGS plotted against a nomogram generated from high PGS CN samples in UKB. (c) equivalent result for low PGS samples from (a). For all sub-figures, the black lines -from top to bottom-represent the 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 97.5% quantiles respectively.

Discussion

We hypothesized that inclusion of genetic information associated with regional brain volume may substantially affect normative models. Indeed, the PGS for HV was significantly positively correlated with estimated HV from MRI; translating into a shift of around 100 mm^3 in nomograms based on PGS stratification (high vs low PGS). Importantly, this magnitude corresponds to ~3 years' worth of HV loss during normal aging for a 65-year-old. While previous studies have examined the impact of disease associated variants, such as *APOE* status, on HV[32,33] our study relied on genetic variants influencing HV in healthy subjects. This is an important difference: the *APOE* genotype is associated with present or future AD status rather than having a direct influence on HV in healthy populations. Indeed, GWAS studies of the hippocampus that exclude dementia patients find little influence of AD associated SNPs[15]. By design, nomograms are intended to model healthy progression and to simplify spotting disease related outliers. Therefore, in theory, accounting for the genetics of healthy variation in HV should enable earlier detection of AD-related HV decline aging individuals. Conversely, stratifying by *APOE*-e4 status when creating HV nomograms charts the different HV trajectories among *APOE* genotypes, however, at the same time masks the pathological decline and thus will theoretically decrease the sensitivity to HV decline.

Subjects with extreme PGS account for significant amounts of the variance as indicated by the S-shape in the quantile plots (e.g., Figure 3). This has been observed in other PGS-trait combinations[19,20,36]. Furthermore, we found similar R^2 values across all PGSs (±0.05 R^2) with two peaks at thresholds of 1E-7 and 0.75. This reflects two types of genetic effects: the first is that few SNPs account for a substantial portion of the total variance in HV because of their high effect size (oligogenic effect) and the second is the combined effect of all common genetic variants on HV (polygenic effect). This type of effect has been reported in other studies of dementia[37].

In addition to demonstrating the clear effect of genetics on normative models, we have shown GPR to be effective for estimating nomograms. Using a model-based method allows us to generate accurate nomograms across the entire age range of the dataset. In fact, our GPR model can potentially be extended a few years beyond those limits (Figure 2 – Figure Supplement 1). In comparison, the SWA nomograms age range is reduced by 20 years

349   compared to the range of the training because of the required smoothing. Thus, compared

350   to the SWA, we extended the age range forwards by 10 years, bringing it out to 82 years old,

351   which is relevant for AD where most patients display symptoms at around age 65-75[27,38].

352   While some methods like LOESS regression can be used to mitigate this need[39], the GPR's

353   model-based approach does not need smoothing to begin with. However, there is a

354   possibility that our results suffer from edge effects. For example, we suspect that the peak

355   noted in the male nomogram is likely due to under-sampling in the younger participants.

356   We found that building nomograms is data efficient: with the SWA, using around 17% (3000

357   samples) of training samples generated nomograms that were on average only 0.4% of

358   average HV (19 mm^3) different to those generated by the full training set. GPR nomograms,

359   achieved the same level of accuracy with only 5% (900 samples) of the dataset (Figure 2 –

360   Figure Supplement 3).

361

362   Using PGS improves the normative modelling in an independent dataset. In ADNI genetic

363   adjustment reduced the percentile gap between similarly diagnosed subjects with

364   genetically predicted high and low HV. The impact of the PGS adjusted model on CN

365   samples was greater than on MCI or AD samples. Genetic adjustment centred the CN

366   samples closer to the 50th percentile. As the effect of building separate nomograms was to

367   mitigate the impact of genetic variability on HV it was not surprising that this effect did not

368   carry over to MCI and AD subjects, likely because the pathological effect of AD on HV (~804

369   mm^3 or 6.4% volume loss) far exceeds the shift in nomograms achieved with genetic

370   adjustment (~100 mm^3 or 0.8% of mean HV). Other studies have found that annual HV loss

371   in CN subjects was between 0.38% and 1.73%[7,40-43]. Using the nomograms from our work,

372   genetic adjustment corresponds to ~3 years of normal aging for a 65-year-old. However,

373   despite this sizable effect, genetically adjusted nomograms did not provide additional

374   insight into distinguishing MCI subjects that remained stable or converted to AD.

375   Nonetheless, the added precision may prove more useful in early detection of deviation

376   among CN subjects, for instance in detecting subtle hippocampal volume loss in individuals

377   with presymptomatic neurodegeneration.

378

379   While this study has shown the significant impact of PGSs on HV nomograms, we have

380   identified areas for improvement. Integrating the PGSs into the GP models would remove

15

381   the need for stratification and allow for more adjustment precision, however, PGSs are

382   prone to 'site' effects depending on the method and quality of genotyping and imputation.

383   Consequently, using the 'raw' PGSs in predictive models presents its own challenges. Also,

384   the PGSs used in this study were based on a GWAS of average bilateral HV in both male and

385   female participants. Previous studies have shown a significant difference between these

386   groups[30], and nomograms estimated for these separate groups are distinct[28,44,45] (Figure 2).

387   Therefore, using separate GWASs for each of these strata would potentially give the PGSs

388   more accuracy. A second limitation of this study is the reliability of HV estimates. There is a

389   significant difference between manual and automated segmentation of the

390   hippocampus[28,44,45]; more so than other brain regions[46,47], and Freesurfer is known to

391   consistently overestimate HV[48]. Therefore, other brain regions with higher SNP heritability

392   like the cerebellum or whole brain volume[14] may show more sensitivity on nomograms.

393   Moreover, a recent study of PGS uncertainty revealed large variance in PGS estimates[49],

394   which may undermine PGS based stratification; hence a more sophisticated method of

395   building PGS or stratification may improve results further. Finally, while NeuroCombat has

396   been shown to remove most site effects, some may remain and still need to be adjusted for

397   [50].

398

399   In conclusion, our study demonstrated that PGS for HV was significantly positively

400   correlated with HV, translating into a shift in the nomograms corresponding to ~3 years'

401   worth of normal aging HV loss for a 65-year-old. We have additionally shown that this effect

402   can be observed in an independent dataset. And while more work in this direction is

403   needed, successful integration of polygenic effects on multiple brain regions may help

404   improve the sensitivity to detect early disease processes.

405

406

407

408

409

410

411

412

413

414

Materials and Methods

Datasets

Data from a total of 39,664 subjects (18,718 female) aged 44 to 82 were obtained from the

UKB (application number 65299) with available genotyping and imaging data. Imaging and

genetic protocols are described in Bycroft et al. (2018)[51] and Miller et al. (2016) [52],

respectively. Briefly, for this analysis we used hippocampal volumes (HV) estimated with

FreeSurfer[53] at the initial imaging visit. The dataset preparation followed the process

described by Nobis et. (2019)[30]. To ensure nomograms represent the spectrum of healthy

aging, subjects were excluded based on history of neurological or psychiatric disorders,

head trauma, substance abuse, or cardiovascular disorders. Furthermore, to control for

population level genetic heterogeneity, only subjects with 'British' ethnic backgrounds were

considered. The dataset was then stratified by self-reported sex. HV outliers were excluded

using mean absolute deviation (MAD) with a threshold of 5.0. Subjects' intracranial volume

(ICV) was derived by using the volumetric scaling from T1 head image to standard space.

Finally, ICV and scan date were linearly regressed out of the HVs.

For an application dataset we used the Alzheimer's Disease Neuroimaging Initiative (ADNI)

database (adni.loni.usc.edu)[54]. 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). A total of 1001 ADNI subjects (445 male) aged 55 to 95 were

included in this analysis. Imaging and genetic protocols are described by Saykin et al.

(2010)[55] and by Jack et al. (2008)[56], respectively. Briefly, we obtained HVs estimated with

FreeSurfer v5.1. Subjects were excluded based on HV quality scores and based on genetic

ancestry (i.e., restricted to self-reported white non-Hispanic ancestry). As with UKB,

estimated volumes were stratified by sex, and ICV and scan date were regressed out of HV

estimates. Finally, we used NeuroCombat[57] to adjust across ADNI sites and harmonize the

volumes with the UKB Dataset. To do this we modelled 58 batches (UKB data as one batch

445 and 57 ADNI sites as separate batches) and added ICV, sex, and diagnosis (assigning all UKB

446 as Healthy and using the diagnosis columns in ADNI) to retain biological variation.

447 Demographics were obtained from the `ADNIMERGE` table (date accessed: 19-06-2020).

448 Furthermore, we used genotyping data of ADNI subjects pre-processed as previously

449 described by Scelsi et. (2018)[58].

450

451 *Sliding Window Approach*

452 As a baseline, we generated nomograms using the sliding window approach (SWA)

453 described by Nobis et al. (2019)[30]. Briefly, we sorted UKB samples by age, and formed 100

454 quantile bins, each containing 10% percent of the samples. This means that neighbouring

455 bins had a 90% overlap. For example, if we had 5,000 samples, each bin contained 500

456 samples and consecutive bins were shifted by 50 samples. Thus, bin number four would

457 start at index 151. Then, within each bin, the 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and

458 97.5% quantiles were calculated. The quantiles were then smoothed with a gaussian kernel

459 of width 20. The smoothing was needed because towards the ends of the data, the sliding

460 windows approach becomes sensitive to noise.

461

462 *Gaussian Process Regression*

463 Our proposed approach uses GPR to build nomograms. Briefly, a GP is a probability

464 distribution over possible functions that fit a set of points[59,60]. In our application it is a

465 distribution of possible 'HV trajectories across age'. The GPR models were trained with the

466 `laGP` [61] R library, which implements a local approximation method that allows large

467 datasets to be trained on consumer grade machines. We applied the commonly used

468 squared exponential covariance kernel function:

$$K(x_1, x_2) = \sigma^2 e^{-\frac{(x_1 - x_2)^2}{2L^2}},$$

469 where x_1 and x_2 are any two age values from the training set. The kernel function is hyper-

470 parameterized by a vertical scale (σ) and a length scale (L), which, following initialization,

471 are fitted using maximum likelihood estimation. The vertical scale is initialized to the mean

472 HV of all samples, and the length scale is initialized to mean age difference between all

473 samples. We trained models of left, right, and mean HV for each sex. Thanks to their

474 probabilistic formulation, GP models naturally provide a standard deviation from which

475   quantiles can be easily computed. After training, we generated models for ages 45 to 82 by

476   increments of 0.25 years, and quantile curves at 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%,

477   and 97.5%. The UKB dataset was used to train the models and the ADNI dataset was used to

478   test them. For all GPR models, we only tested the ADNI samples that lay within the age

479   range of each model respectively.

480

481

482   *Polygenic Score for Hippocampal Volume*

483   A polygenic score (PGS) is a sum of the impact of a selection of genetic variants on a trait,

484   weighted by the allele count. That is:

$$PGS = \sum_{\forall\, i\, \in\, SNPs} ES_i * C_i,$$

485   where (ES_i) is the effect size (e.g., beta or log(odds) ratio from GWAS summary statistics),

486   and (C_i) is the allele count of SNP i in the subject (either 0,1 or 2). Thus, computing PGSs

487   requires SNP-level genetic data. Using a previously reported GWAS of mean bilateral HV

488   using 26,814 (European) subjects from the ENIGMA study[15], we built a PGS for HV with

489   PRSice v2[62]. For both UKB and ADNI, we filter for minor allele frequency of 0.05, genotype

490   missingness of 0.1, and clumping at 250kb; after which we were left with 70,251 potential

491   SNPS to include for UKB and 114,812 for ADNI. The most widely applied strategy for SNP

492   selection is p-value thresholding. We generated PGSs at 14 p-value thresholds (1E-8, 1E-7,

493   1E-6, 1E-5, 1E-4, 1E-3, 0.01, 0.05, 0.1, 0.2, 0.4, 0.5, 0.75, 1). These thresholds produced a

494   range of polygenic scores comprising as little as 6 SNPs (p-value cut-off at 1E-8) to all

495   available SNPs (p-value cut-off at 1.0). Model fit is then checked by regressing HV against

496   these PGSs while accounting for age, age^2, sex, ICV, and ten genetic principal components.

497

498   *Genetically Adjusted Nomograms*

499   Given the high heritability of HV we investigated whether nomograms can be genetically

500   adjusted. Specifically, we used the top and bottom 30% samples by PGS (at p-value < 0.75

501   threshold) separately to build genetically adjusted nomograms. We found that using a 30%

502   cut-off provided a balance of training size and performance (Figure 2 – Figure Supplement

503   4). Thus, PGS provided us with a way to place new samples in their 'appropriate' nomogram.

504   For instance, within the ADNI dataset we generated PGSs and split the top and bottom (i.e.,

505   high, and low expected HV, respectively) to test against genetically adjusted UKB

506   nomograms. To evaluate the impact of genetic adjustment, we perform a series of ANOVA

507   tests across adjusted nomograms. E.g., we performed an ANOVA test of the HV percentiles

508   of the top 30% UKB samples in the unadjusted then the adjusted nomograms. We did the

509   same for bottom 30% and for men and women. To assess the specificity of the HV-based

510   PGS, we performed this genetic adjustment using PGSs of ICV and AD based on previously

511   reported GWASs[63,64].

512

513   *Longitudinal Analysis*

514   As nomograms are often used to track progression, we examined the impact of the

515   genetically adjusted nomograms on prospective longitudinal data. To this end, we analysed

516   patients from the ADNI cohort that were initially diagnosed as MCI and either converted to

517   AD (progressor) or remained MCI (stable) within five years of follow-up. We tested whether

518   the PGS-adjusted nomograms improved the separation between stable and progressor

519   patients using Cox proportional hazards models while accounting for sex and age.

520

521   **Code and Data Availability**

522   The scripts and code used in this study have been made publicly available and can be found

523   at: https://github.com/Mo-Janahi/NOMOGRAMS. All underlying data, and derived

524   quantities, are available by application from the UK Biobank at

525   http://www.ukbiobank.ac.uk, and by application from ADNI at

526   http://adni.loni.usc.edu/data-samples/access-data/ . Summary statistics from all genome-

527   wide association studies described in this paper are available from the NHGRI-EBI GWAS

528   catalog, study numbers: GCST003834, GCST002245, and GCST003961. URL:

529   https://www.ebi.ac.uk/gwas/studies/

530

531   **Ethics Statement**

532   This work uses exclusively pseudonymized data that were collected as part of different

533   studies (ADNI and UKB). Research ethics approval was granted by UCL as part of project

534   #13083/002.

535

568   Therapeutic Research Institute at the University of Southern California. ADNI data are

569   disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

570   As such, the investigators within the ADNI contributed to the design and implementation of

571   ADNI and/or provided data but did not participate in analysis or writing of this report.

572   **Supplementary Figure Legends:**

573

574   **Figure 2 – Figure Supplement 1: Expanded GPR Nomogram.** a GPR model trained with
575   mean bilateral HV of male subjects in the age range 45-82 (grey circles) and then generated
576   nomograms for the age range 30-100. Within training data range nomogram follows data
577   reasonably well. Outside data rage, nomogram flairs out from expected range after 2-6
578   years. Fairing is faster in the lower ages because, outside the data range, the GPR model
579   reverts to a normal distribution with zero mean. For all sub-figures, the black lines -from top
580   to bottom-represent the 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 97.5% quantiles
581   respectively.

582

583   **Figure 2 – Figure Supplement 2: Model fit of healthy ADNI subjects.** Nomograms produced
584   from healthy subjects in the UKB using the gaussian process regression (GPR) method.
585   Overlayed are scatter plots Cognitively Normal subjects from the ADNI dataset. Male
586   subjects averaged (56.9% ± 24.6 SD) and Female subjects averaged (54.9 ± 26.5 SD). For
587   both sub-figures, the black lines -from top to bottom-represent the 2.5%, 5%, 10%, 25%,
588   50%, 75%, 90%, 95%, and 97.5% quantiles respectively.

589   **Figure 2 – Figure Supplement 3: Performance of GPR and SWM across sample size.** Model
590   training progression is shown for both SWM (top row) and GPR (middle row) models at
591   representative training sizes. Performance (bottom figure) is summarized using the mean
592   distance between generated nomograms and the GPR nomograms built with the full
593   training set (~15k) (shaded areas in the top two rows). By repeatedly sampling data from
594   across age (10 times at each training sample size), we plot the average performance and
595   95% CI of each method. Both methods are data efficient, SWM can achieve 20 mm3 mean
596   difference (0.4% of mean HV) performance using ~3000 samples (20% of training set), and
597   GPR can achieve the same performance using only 1000 samples (~7% of training set).

598   **Figure 2 – Figure Supplement 4: GPR model across top/bottom thresholds.** Illustrated is
599   Male bilateral HV. When stratifying by PRS, there is a trade-off between training set size and
600   final model performance. In these figures, performance is measured by average distance
601   between the percentile curves. At 10%, (leftmost column), the top/bottom strata contain
602   ~1500 samples each and the mean distance is 65 mm3, and at 50% (the rightmost column)
603   they contain ~7500 and the mean distance is 21.5 mm3. For all sub-figures, the black lines -
604   from top to bottom-represent the 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 97.5%
605   quantiles respectively.

606

607   **Figure 3 – Figure Supplement 1: Summary of PGSs based on HV GWAS in UKB samples.** The
608   left set of graphs show the R-Squared of the regression models of PRS across HV for the

609 scores built across SNP P-Value thresholds. While the difference is small, we consistently see
610 a dip in the R-Squared for the middle set of thresholds. The set of figures to the right show
611 the spread of HV across PRS percentile. We display the percentiles for the 0.75 threshold as
612 it showed the best correlation with HV overall.

613
614 **Figure 3 – Figure Supplement 2: Summary of PGSs and models based on HV GWAS and**
615 **ADNI samples.** The left set of graphs show the R-Squared of the regression models of PRS
616 across HV for the scores built across SNP P-Value thresholds. In contrast to the graphs seen
617 in the UKB samples, the R-squared values for the most part increase with p-value threshold.
618 The set of graphs on the right show the spread of HV across PGS percentiles, each at the
619 score that had the highest R-Squared value from the corresponding left graph.
620

621 **Figure 4 – Figure Supplement 1: Genetically Adjusted Nomograms.** For all sub-figures, the
622 black lines -from top to bottom-represent the 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%,
623 and 97.5% quantiles respectively.

624 **Figure 4 – Figure Supplement 2**: Nomograms generated with the SWM by stratifying the
625 sample set based on PGSs. Left column: PGS based on HV GWAS. Middle column: PGS based
626 on ICV GWAS. Right column: PGS based on AD GWAS. For all sub-figures, the black lines -
627 from top to bottom-represent the 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 97.5%
628 quantiles respectively.

629 **Figure 4 – Figure Supplement 3: Training Data Ridge Plots.** Histograms of bilateral HV
630 across the different subsets of the datasets. Samples are grouped in bins of 5 years. N is the
631 number of samples in each set and p is the p-value from a Shapiro-Wilks test of normality.
632 Typically, this test would indicate a non-gaussian distribution with a p-value lower than 0.05
633 (0.001 corrected for 48 multiple tests in this case).
634
635
636
637 # References
638
639 1.   Shen L, Thompson PM. Brain Imaging Genomics: Integrated Analysis and Machine
640 Learning. *Proceedings of the IEEE*. 2020;108(1):125-162. doi:10.1109/JPROC.2019.2947272
641 2.   Marquand AF, Rezek I, Buitelaar J, Beckmann CF. Understanding Heterogeneity in
642 Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biological*
643 *Psychiatry*. 2016;80(7):552-561. doi:10.1016/j.biopsych.2015.12.023
644 3.   Pinaya WHL, Scarpazza C, Garcia-Dias R, et al. Normative modelling using deep
645 autoencoders: A multi-cohort study on mild cognitive impairment and Alzheimer's disease.
646 bioRxiv; 2020. p. 2020.02.10.931824-2020.02.10.931824.
647 4.   Wolfers T, Beckmann CF, Hoogman M, Buitelaar JK, Franke B, Marquand AF.
648 Individual differences v. the average patient: Mapping the heterogeneity in ADHD using
649 normative models. *Psychological Medicine*. 2019;50(2):314-323.
650 doi:10.1017/S0033291719000084

651   5.   Ziegler G, Ridgway GR, Dahnke R, Gaser C. Individualized Gaussian process-based
652 prediction and detection of local and global gray matter abnormalities in elderly subjects.
653 *NeuroImage*. 2014;97:333-348. doi:10.1016/j.neuroimage.2014.04.018

654   6.   Xavier Castellanos F, Lee PP, Sharp W, et al. Developmental trajectories of brain
655 volume abnormalities in children and adolescents with attention-deficit/hyperactivity
656 disorder. *Journal of the American Medical Association*. 2002;288(14):1740-1748.
657 doi:10.1001/jama.288.14.1740

658   7.   Scahill RI, Frost C, Jenkins R, Whitwell JL, Rossor MN, Fox NC. A longitudinal study of
659 brain volume changes in normal aging using serial registered magnetic resonance imaging.
660 *Archives of Neurology*. 2003;60(7):989-994. doi:10.1001/archneur.60.7.989

661   8.   Peterson M, Warf BC, Schiff SJ. Normative human brain volume growth. *Journal of
662 Neurosurgery: Pediatrics*. 2018;21(5):478-485. doi:10.3171/2017.10.PEDS17141

663   9.   Lukies MW, Watanabe Y, Tanaka H, et al. Heritability of brain volume on MRI in
664 middle to advanced age: A twin study of Japanese adults. *PLoS ONE*.
665 2017;12(4)doi:10.1371/journal.pone.0175800

666   10.   Kremen WS, Prom-Wormley E, Panizzon MS, et al. Genetic and environmental
667 influences on the size of specific brain regions in midlife: The VETSA MRI study. *NeuroImage*.
668 2010-01-01 2010;49(2):1213-1223. doi:10.1016/j.neuroimage.2009.09.043

669   11.   Thompson PM, Jahanshad N, Ching CRK, et al. ENIGMA and global neuroscience: A
670 decade of large-scale studies of the brain in health and disease across more than 40
671 countries. Springer Nature; 2020. p. 1-28.

672   12.   Hibar DP, Stein JL, Renteria ME, et al. Common genetic variants influence human
673 subcortical brain structures. *Nature*. 2015-04-01 2015;520(7546):224-229.
674 doi:10.1038/nature14101

675   13.   Rentería ME, Hansell NK, Strike LT, et al. Genetic architecture of subcortical brain
676 regions: Common and region-specific genetic contributions. *Genes, Brain and Behavior*.
677 2014;13(8):821-830. doi:10.1111/gbb.12177

678   14.   Zhao B, Ibrahim JG, Li Y, et al. Heritability of Regional Brain Volumes in Large-Scale
679 Neuroimaging and Genetic Studies. *Cerebral Cortex*. 2019;29(7):2904-2914.
680 doi:10.1093/cercor/bhy157

681   15.   Hibar DP, Adams HHH, Jahanshad N, et al. Novel genetic loci associated with
682 hippocampal volume. *Nature Communications*. 2017;8doi:10.1038/ncomms13624

683   16.   Grasby KL, Jahanshad N, Painter JN, et al. The genetic architecture of the human
684 cerebral cortex. *Science*. 2020-03-20 2020;367(6484):eaay6690.
685 doi:10.1126/science.aay6690

686   17.   Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex
687 diseases. *Nature*. 2009-10-01 2009;461(7265):747-753. doi:10.1038/nature08494

688   18.   Foo H, Thalamuthu A, Jiang J, et al. Associations between Alzheimer's disease
689 polygenic risk scores and hippocampal subfield volumes in 17,161 UK Biobank participants.
690 *Neurobiology of Aging*. 2021;98:108-115. doi:10.1016/j.neurobiolaging.2020.11.002

691   19.   Axelrud LK, Santoro ML, Pine DS, et al. Polygenic risk score for Alzheimer's disease:
692 Implications for memory performance and hippocampal volumes in early life. *American
693 Journal of Psychiatry*. 2018;175(6):555-563. doi:10.1176/appi.ajp.2017.17050529

694   20.   Escott-Price V, Nalls MA, Morris HR, et al. Polygenic risk of <scp>P</scp> arkinson
695 disease is correlated with disease age at onset. *Annals of Neurology*. 2015;77(4):582-591.
696 doi:10.1002/ana.24335

697 21. Bird CM, Burgess N. The hippocampus and memory: Insights from spatial processing.
698 Nature Publishing Group; 2008. p. 182-194.
699 22. Bremner JD, Narayan M, Anderson ER, Staib LH, Miller HL, Charney DS. Hippocampal
700 Volume Reduction in Major Depression. *American Journal of Psychiatry*. 2000-01-01
701 2000;157(1):115-118. doi:10.1176/ajp.157.1.115
702 23. Nelson MD, Saykin AJ, Flashman LA, Riordan HJ. Hippocampal Volume Reduction in
703 Schizophrenia as Assessed by Magnetic Resonance Imaging. *Archives of General Psychiatry*.
704 1998-05-01 1998;55(5):433. doi:10.1001/archpsyc.55.5.433
705 24. Whelan CD, Altmann A, Botía JA, et al. Structural brain abnormalities in the common
706 epilepsies assessed in a worldwide ENIGMA study. *Brain : a journal of neurology*.
707 2018;141(2):391-408. doi:10.1093/brain/awx341
708 25. Pini L, Pievani M, Bocchetta M, et al. Brain atrophy in Alzheimer's Disease and aging.
709 *Ageing Research Reviews*. 2016-09-01 2016;30:25-48. doi:10.1016/j.arr.2016.01.002
710 26. Van Der Flier WM, Scheltens P. Epidemiology and risk factors of dementia. BMJ
711 Publishing Group Ltd; 2005. p. 2-7.
712 27. Rabinovici GD. Late-onset Alzheimer disease. Lippincott Williams and Wilkins; 2019.
713 p. 14-33.
714 28. Schmidt MF, Storrs JM, Freeman KB, et al. A comparison of manual tracing and
715 FreeSurfer for estimating hippocampal volume over the adult lifespan. *Human Brain
716 Mapping*. 2018;39(6):2500-2513. doi:10.1002/hbm.24017
717 29. Fraser MA, Shaw ME, Cherbuin N. A systematic review and meta-analysis of
718 longitudinal hippocampal atrophy in healthy human ageing. Academic Press Inc.; 2015. p.
719 364-374.
720 30. Nobis L, Manohar SG, Smith SM, et al. Hippocampal volume across age: Nomograms
721 derived from over 19,700 people in UK Biobank. *NeuroImage: Clinical*.
722 2019;23doi:10.1016/j.nicl.2019.101904
723 31. Mather KA, Armstrong NJ, Wen W, et al. Investigating the Genetics of Hippocampal
724 Volume in Older Adults without Dementia. *PLOS ONE*. 2015-01-27 2015;10(1):e0116920.
725 doi:10.1371/journal.pone.0116920
726 32. Ching C, Abaryan Z, Santhalingam V, et al. Sex differences in subcortical aging: A
727 nomogram study of age, sex, and apoe (N = 9,414). *Alzheimer's & Dementia*.
728 2020;16(S4):e045774-e045774. doi:10.1002/alz.045774
729 33. Veldsman M, Nobis L, Alfaro-Almagro F, Manohar S, Husain M. The human
730 hippocampus and its subfield volumes across age, sex and APOE e4 status. *Brain
731 Communications*. 2021;3(1)doi:10.1093/braincomms/fcaa219
732 34. Kim J, Basak JM, Holtzman DM. The Role of Apolipoprotein E in Alzheimer's Disease.
733 NIH Public Access; 2009. p. 287-303.
734 35. Liu CC, Kanekiyo T, Xu H, Bu G. Apolipoprotein e and Alzheimer disease: Risk,
735 mechanisms and therapy. NIH Public Access; 2013. p. 106-118.
736 36. Ranlund S, Calafato S, Thygesen JH, et al. A polygenic risk score analysis of psychosis
737 endophenotypes across brain functional, structural, and cognitive domains. *American
738 Journal of Medical Genetics Part B: Neuropsychiatric Genetics*. 2018;177(1):21-34.
739 doi:10.1002/ajmg.b.32581
740 37. Bis JC, Decarli C, Smith AV, et al. Common variants at 12q14 and 12q24 are
741 associated with hippocampal volume. *Nature Genetics*. 2012;44(5):545-551.
742 doi:10.1038/ng.2237
743 38. Mendez MF. Early-Onset Alzheimer Disease. W.B. Saunders; 2017. p. 263-281.

744 39.   Bethlehem RAI, Seidlitz J, Romero-Garcia R, Trakoshis S, Dumas G, Lombardo MV. A
745 normative modelling approach reveals age-atypical cortical thickness in a subgroup of males
746 with autism spectrum disorder. *Communications Biology*. 2020-12-01
747 2020;3(1)doi:10.1038/s42003-020-01212-9
748 40.   Leong RLF, Lo JC, Sim SKY, et al. Longitudinal brain structure and cognitive changes
749 over 8 years in an East Asian cohort. *NeuroImage*. 2017;147:852-860.
750 doi:10.1016/j.neuroimage.2016.10.016
751 41.   Jack CR, Petersen RC, Xu Y, et al. Rates of hippocampal atrophy correlate with
752 change in clinical status in aging and AD. *Neurology*. 2000;55(4):484-489.
753 doi:10.1212/wnl.55.4.484
754 42.   Mori E, Lee K, Yasuda M, et al. Accelerated hippocampal atrophy in Alzheimer's
755 disease with apolipoprotein E ?4 allele. *Annals of Neurology*. 2002-02-01 2002;51(2):209-
756 214. doi:10.1002/ana.10093
757 43.   Risacher SL, Shen L, West JD, et al. Longitudinal MRI atrophy biomarkers:
758 Relationship to conversion in the ADNI cohort. *Neurobiology of Aging*. 2010-08-01
759 2010;31(8):1401-1418. doi:10.1016/j.neurobiolaging.2010.04.029
760 44.   Khlif MS, Egorova N, Werden E, et al. A comparison of automated segmentation and
761 manual tracing in estimating hippocampal volume in ischemic stroke and healthy control
762 participants. *NeuroImage: Clinical*. 2019;21:101581-101581. doi:10.1016/j.nicl.2018.10.019
763 45.   Pardoe HR, Pell GS, Abbott DF, Jackson GD. Hippocampal volume assessment in
764 temporal lobe epilepsy: How good is automated segmentation? *Epilepsia*.
765 2009;50(12):2586-2592. doi:10.1111/j.1528-1167.2009.02243.x
766 46.   Keller SS, Gerdes JS, Mohammadi S, et al. Volume estimation of the thalamus using
767 FreeSurfer and stereology: Consistency between methods. *Neuroinformatics*.
768 2012;10(4):341-350. doi:10.1007/s12021-012-9147-0
769 47.   Buser NJ, Madan CR, Hanson JL. Quantifying numerical and spatial reliability of
770 amygdala and hippocampal subdivisions in FreeSurfer. bioRxiv; 2020. p. 2020.06.12.149203-
771 2020.06.12.149203.
772 48.   Perlaki G, Horvath R, Nagy SA, et al. Comparison of accuracy between FSL's FIRST and
773 Freesurfer for caudate nucleus and putamen segmentation. *Scientific Reports*. 2017;7(1):1-
774 9. doi:10.1038/s41598-017-02584-5
775 49.   Ding Y, Hou K, Burch KS, et al. Large uncertainty in individual PRS estimation impacts
776 PRS-based risk stratification. bioRxiv; 2020. p. 2020.11.30.403188-2020.11.30.403188.
777 50.   Stamoulou E, Manikis GC, Tsiknakis M, Marias K. ComBat harmonization for
778 multicenter MRI based radiomics features. IEEE; 2021:
779 51.   Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep
780 phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-
781 0579-z
782 52.   Miller KL, Alfaro-Almagro F, Bangerter NK, et al. Multimodal population brain
783 imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*.
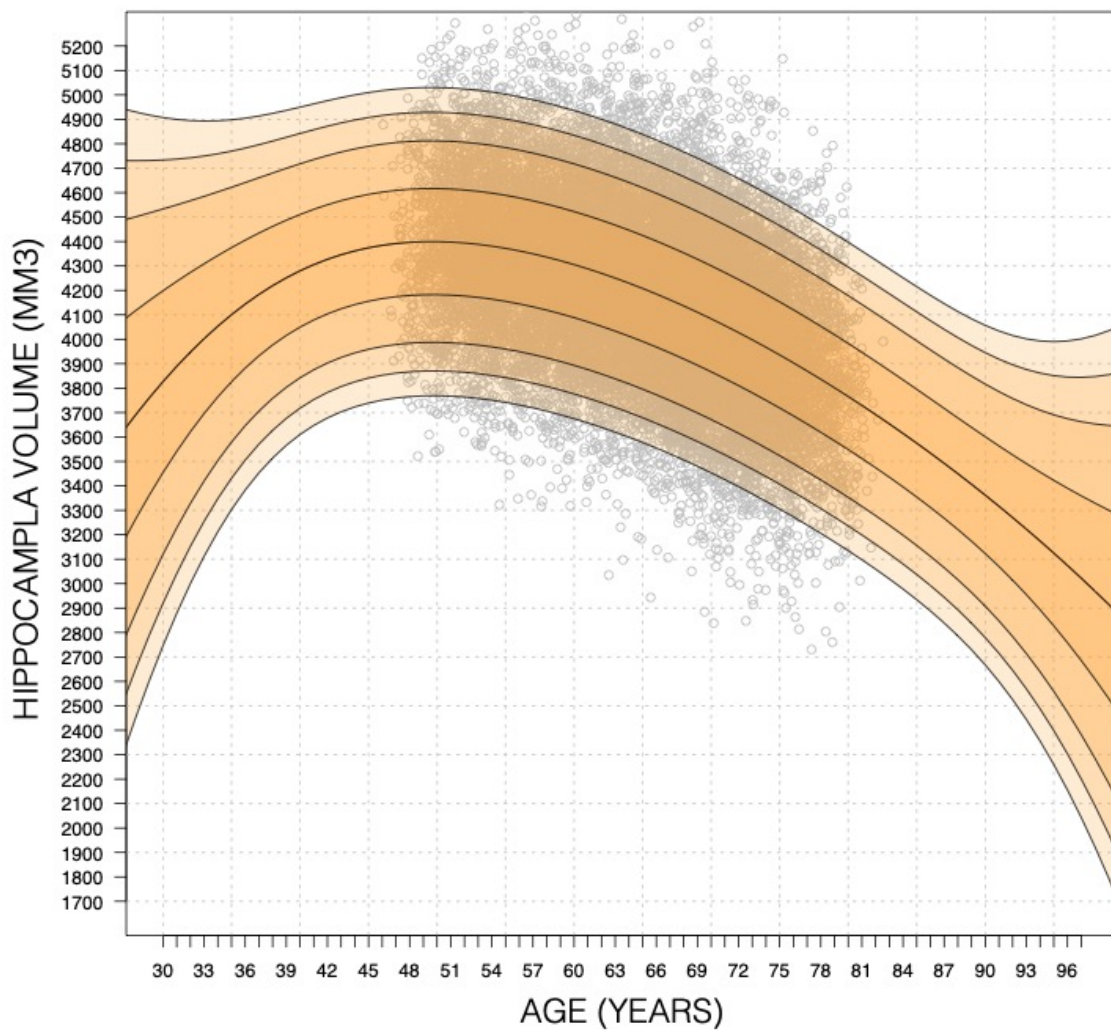784 2016;19(11):1523-1536. doi:10.1038/nn.4393
785 53.   Fischl B. FreeSurfer. NIH Public Access; 2012. p. 774-781.
786 54.   Petersen RC, Aisen PS, Beckett LA, et al. Alzheimer's Disease Neuroimaging Initiative
787 (ADNI): Clinical characterization. *Neurology*. 2010;74(3):201-209.
788 doi:10.1212/WNL.0b013e3181cb3e25

789   55.   Saykin AJ, Shen L, Foroud TM, et al. Alzheimer's Disease Neuroimaging Initiative
790   biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. *Alzheimer's*
791   *and Dementia*. 2010;6(3):265-273. doi:10.1016/j.jalz.2010.03.013
792   56.   Jack CR, Bernstein MA, Fox NC, et al. The Alzheimer's disease neuroimaging initiative
793   (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*. 2008;27(4):685-691.
794   doi:10.1002/jmri.21049
795   57.   Fortin JP, Cullen N, Sheline YI, et al. Harmonization of cortical thickness
796   measurements across scanners and sites. *NeuroImage*. 2018;167:104-120.
797   doi:10.1016/j.neuroimage.2017.11.024
798   58.   Scelsi MA, Khan RR, Lorenzi M, et al. Genetic study of multimodal imaging
799   Alzheimer's disease progression score implicates novel loci. *Brain*. 2018;141(7):2167-2180.
800   doi:10.1093/brain/awy141
801   59.   Rasmussen CE, Williams CKI. *Gaussian Processes for Machine Learning*. 2006.
802   026218253X. www.GaussianProcess.org/gpml
803   60.   Wang J. *An Intuitive Tutorial to Gaussian Processes Regression*. 2021.
804   61.   Gramacy RB. LaGP: Large-scale spatial modeling via local approximate Gaussian
805   processes in R. *Journal of Statistical Software*. 2016;72(1):1-46. doi:10.18637/jss.v072.i01
806   62.   Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data.
807   *GigaScience*. 2019;8(7):1-6. doi:10.1093/gigascience/giz082
808   63.   Adams HHH, Hibar DP, Chouraki V, et al. Novel genetic loci underlying human
809   intracranial volume identified through genome-wide association. *Nature Neuroscience*.
810   2016;19(12):1569-1582. doi:10.1038/nn.4398
811   64.   Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals
812   identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*.
813   2013;45(12):1452-1458. doi:10.1038/ng.2802
814
815
816
817
818
819
820
821
822
823
824
825
826
827

828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874

Supplementary Figures

Figure 2 – Figure Supplement 1: Expanded GPR Nomogram. a GPR model trained with mean bilateral HV of male subjects in the age range 45-82 (grey circles) and then generated nomograms for the age range 30-100. Within training data range nomogram follows data reasonably well. Outside data rage, nomogram flairs out from expected range after 2-6 years. Fairing is faster in the lower ages because, outside the data range, the GPR model reverts to a normal distribution with zero mean. For all sub-figures, the black lines -from top to bottom-represent the 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 97.5% quantiles respectively.

875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
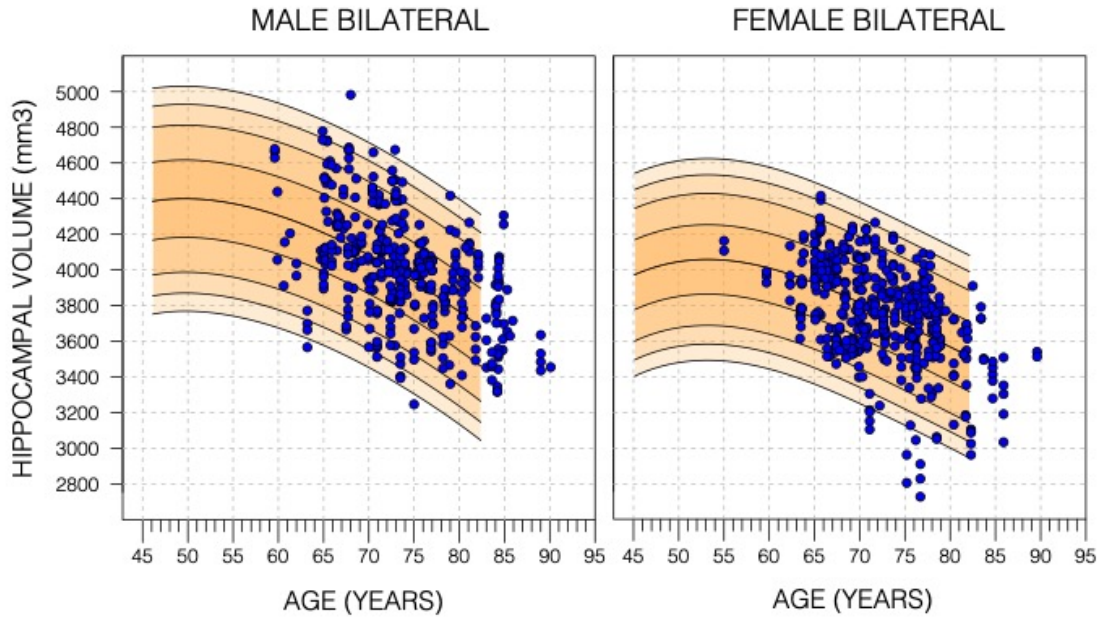
Figure 2 – Figure Supplement 2: Model fit of healthy ADNI subjects. Nomograms produced from healthy subjects in the UKB using the gaussian process regression (GPR) method. Overlayed are scatter plots Cognitively Normal subjects from the ADNI dataset. Male subjects averaged (56.9% ± 24.6 SD) and Female subjects averaged (54.9 ± 26.5 SD). For both sub-figures, the black lines -from top to bottom-represent the 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 97.5% quantiles respectively.
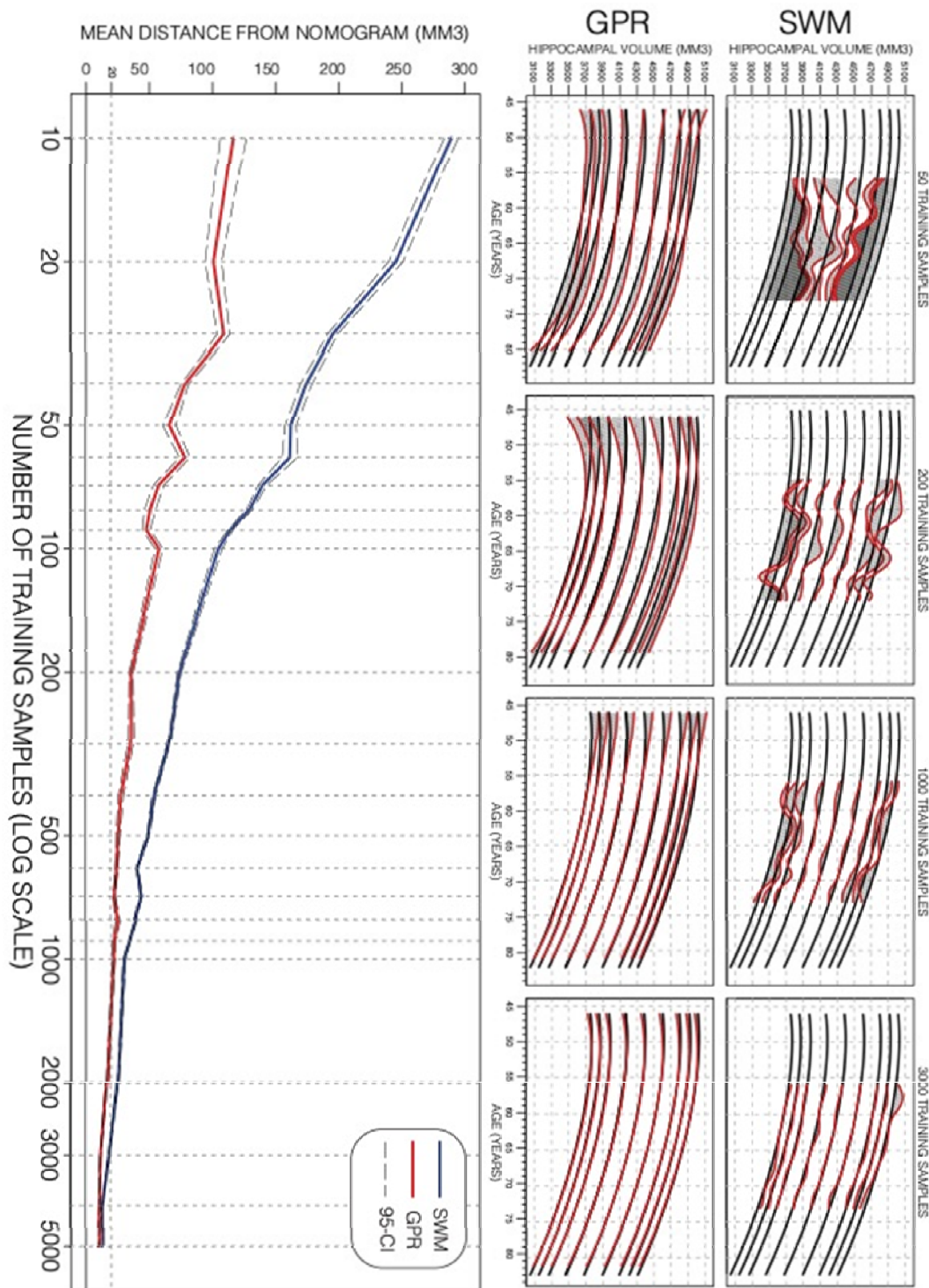
959 **Figure 2 – Figure Supplement 3: Performance of GPR and SWM across sample size.** Model
960 training progression is shown for both SWM (top row) and GPR (middle row) models at
961 representative training sizes. Performance (bottom figure) is summarized using the mean
962 distance between generated nomograms and the GPR nomograms built with the full
963 training set (~15k) (shaded areas in the top two rows). By repeatedly sampling data from
964 across age (10 times at each training sample size), we plot the average performance and
965 95% CI of each method. Both methods are data efficient, SWM can achieve 20 mm3 mean
966 difference (0.4% of mean HV) performance using ~3000 samples (20% of training set), and
967 GPR can achieve the same performance using only 1000 samples (~7% of training set).

968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
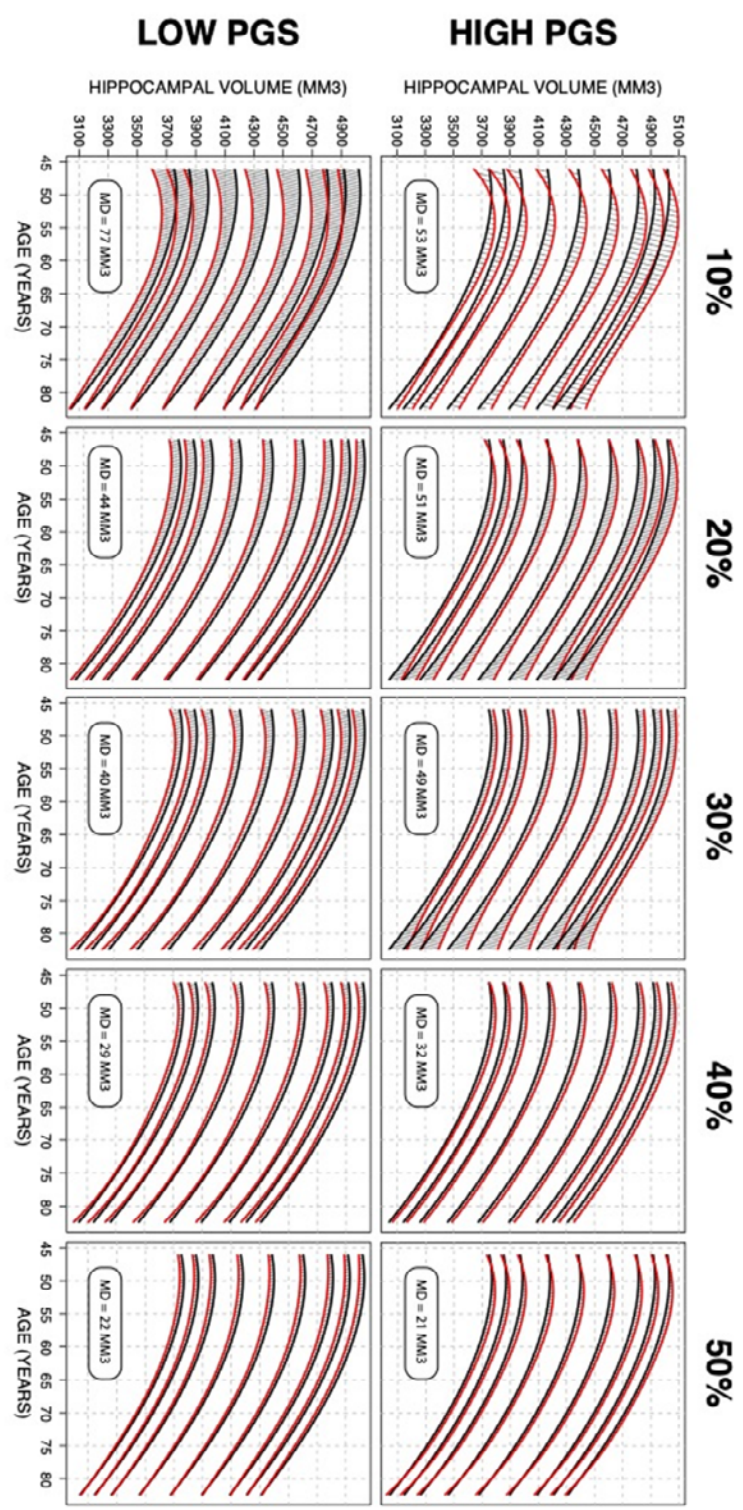
1005 **Figure 2 – Figure Supplement 4: GPR model across top/bottom thresholds.** Illustrated is
1006 Male bilateral HV. When stratifying by PRS, there is a trade-off between training set size and
1007 final model performance. In these figures, performance is measured by average distance
1008 between the percentile curves. At 10%, (leftmost column), the top/bottom strata contain
1009 ~1500 samples each and the mean distance is 65 mm3, and at 50% (the rightmost column)
1010 they contain ~7500 and the mean distance is 21.5 mm3. For all sub-figures, the black lines -
1011 from top to bottom-represent the 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 97.5%
1012 quantiles respectively.

1013
1014

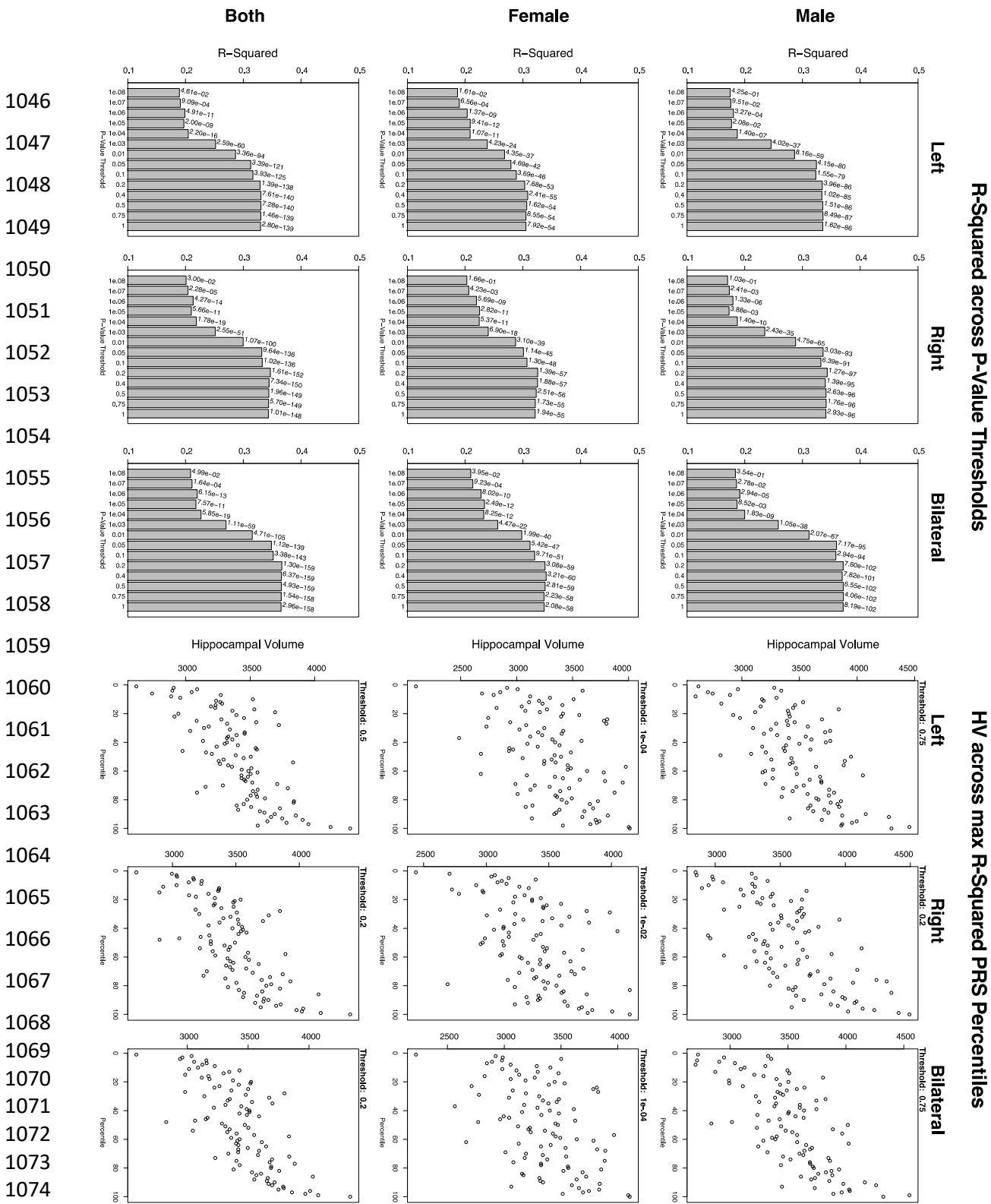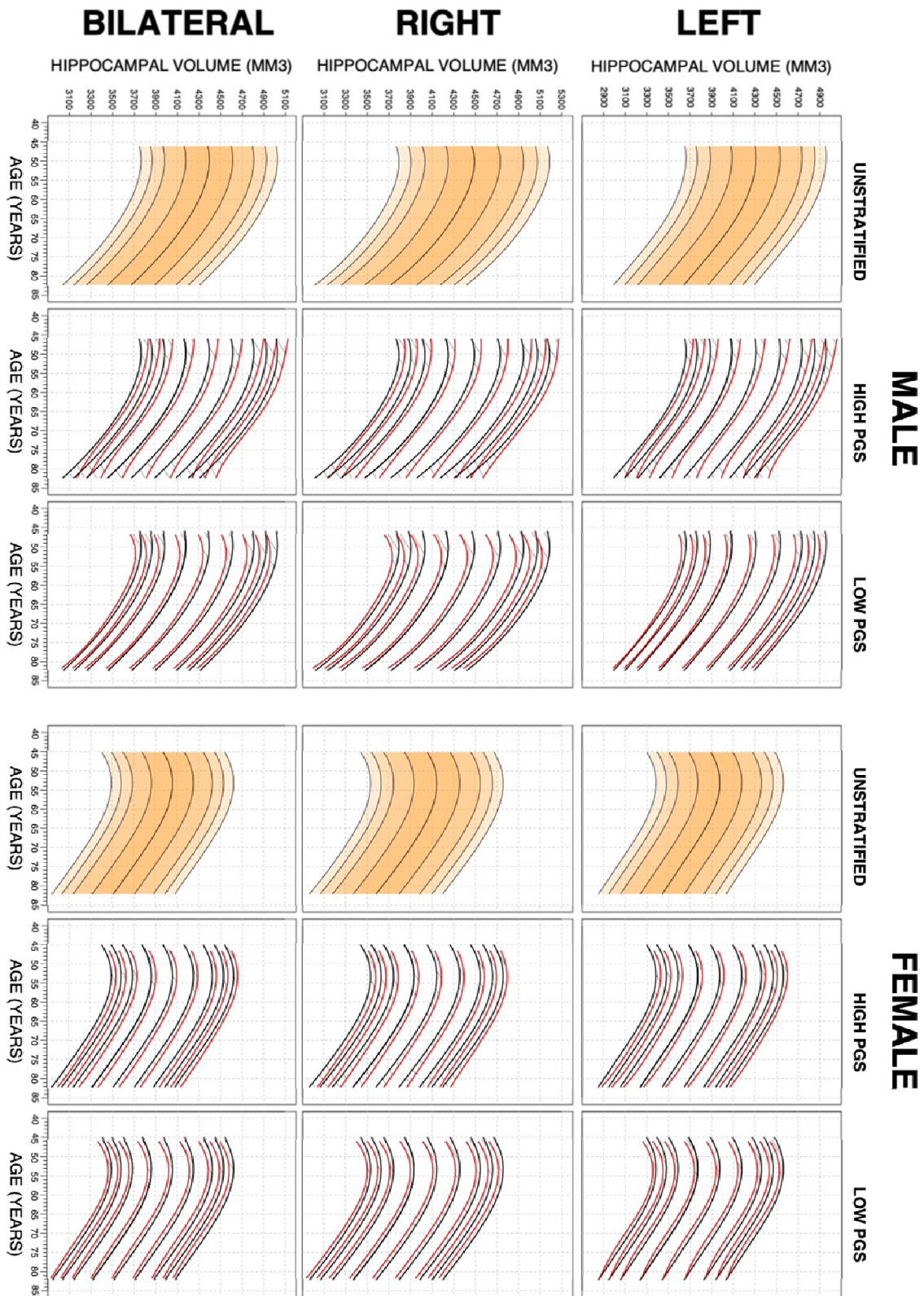**Figure 3 – Figure Supplement 1: Summary of PGSs based on HV GWAS in UKB samples.** The left set of graphs show the R-Squared of the regression models of PRS across HV for the scores built across SNP P-Value thresholds. While the difference is small, we consistently see a dip in the R-Squared for the middle set of thresholds. The set of figures to the right show the spread of HV across PRS percentile. We display the percentiles for the 0.75 threshold as it showed the best correlation with HV overall.

Both   Female   Male

R-Squared across P-Value Thresholds

HV across max R-Squared PRS Percentiles

Figure 3 – Figure Supplement 2: Summary of PGSs and models based on HV GWAS and
ADNI samples. The left set of graphs show the R-Squared of the regression models of PRS
across HV for the scores built across SNP P-Value thresholds. In contrast to the graphs seen
in the UKB samples, the R-squared values for the most part increase with p-value threshold.
The set of graphs on the right show the spread of HV across PGS percentiles, each at the
score that had the highest R-Squared value from the corresponding left graph.

33

Figure 4 – Figure Supplement 1: Genetically Adjusted Nomograms. For all sub-figures, the black lines -from top to bottom-represent the 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 97.5% quantiles respectively.

Figure 4 – Figure Supplement 2: Nomograms generated with the SWM by stratifying the sample set based on PGSs. Left column: PGS based on HV GWAS. Middle column: PGS based on ICV GWAS. Right column: PGS based on AD GWAS. For all sub-figures, the black lines - from top to bottom-represent the 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 97.5% quantiles respectively.

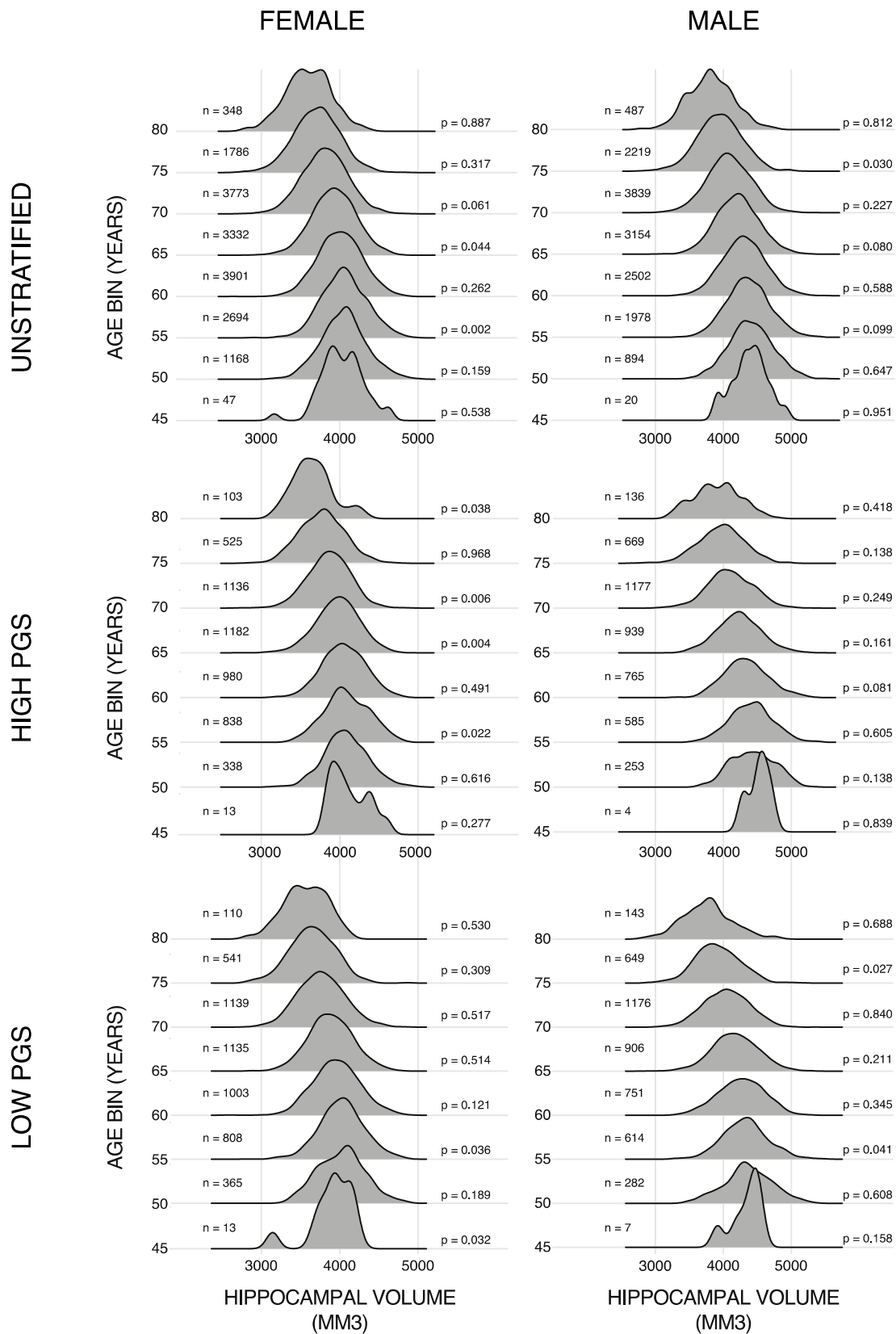FEMALE   MALE

Figure 4 – Figure Supplement 3: Training Data Ridge Plots. Histograms of bilateral HV across the different subsets of the datasets. Samples are grouped in bins of 5 years. N is the number of samples in each set and p is the p-value from a Shapiro-Wilks (SW) test of normality. Typically, this test would indicate a non-gaussian distribution with a p-value lower than 0.05 (0.001 corrected for 48 multiple tests in this case).