RESEARCH ARTICLE

Dementias Platform UK: Bringing genetics into life

Ganna Leonenko[1] [iD] | **Sarah Bauermeister**[2] | **Dipanwita Ghanti**[2] |
Joshua Stevenson-Hoare[3] | **Emily Simmonds**[1] | **Keeley Brookes**[4] | **Kevin Morgan**[5] |
Nishi Chaturvedi[6] | **Paul Elliott**[7,8] | **Alan Thomas**[9] | **Nicholas Wareham**[10] |
John Gallacher[2] | **Valentina Escott-Price**[1,3] [iD]

[1]Dementia Research Institute, Cardiff University, Cardiff, UK

[2]Dementias Platform UK, Department of Psychiatry, University of Oxford, Oxford, UK

[3]Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK

[4]Biosciences, School of Science and Technology, Nottingham Trent University, Nottingham, UK

[5]University Nottingham, Nottingham, UK

[6]MRC Unit for Lifelong Health & Ageing, UCL, London, UK

[7]MRC Centre for Environment and Health, School of Public Health, Imperial College London, London, UK

[8]UK Dementia Research Institute, Imperial College London, London, UK

[9]Translational and Clinical Research Institute, Newcastle University, Newcastle, UK

[10]MRC Epidemiology Unit, University of Cambridge, Cambridge, UK

Correspondence
Valentina Escott-Price, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Dementia Research Institute, Cardiff University, Cardiff CF24 4HQ, UK.
Email: escottpricev@cardiff.ac.uk

John Gallacher, Dementias Platform UK, Department of Psychiatry, University of Oxford, Oxford OX3 7JX, UK.
Email: john.gallacher@psych.ox.ac.uk

Ganna Leonenko and Sarah Bauermeister are joint first authors.

Abstract

INTRODUCTION: The Dementias Platform UK (DPUK) Data Portal is a data repository bringing together a wide range of cohorts. Neurodegenerative dementias are a group of diseases with highly heterogeneous pathology and an overlapping genetic component that is poorly understood. The DPUK collection of independent cohorts can facilitate research in neurodegeneration by combining their genetic and phenotypic data.

METHODS: For genetic data processing, pipelines were generated to perform quality control analysis, genetic imputation, and polygenic risk score (PRS) derivation with six genome-wide association studies of neurodegenerative diseases. Pipelines were applied to five cohorts.

DISCUSSION: The data processing pipelines, research-ready imputed genetic data, and PRS scores are now available on the DPUK platform and can be accessed upon request though the DPUK application process. Harmonizing genome-wide data for multiple datasets increases scientific opportunity and allows the wider research community to access and process data at scale and pace.

KEYWORDS
dementia, genetic data, harmonization, imputation, polygenic risk scores

1 | BACKGROUND

Dementias Platform UK (DPUK; https://www.dementiasplatform.uk/) brings together a wide range of cohorts in the DPUK Data Portal to facilitate collaborative research opportunities and answer important questions about dementia.[1] DPUK is fully auditable with a remote access platform that contains > 60 population and clinical cohorts across a range of imaging, genetic, and survey (e.g., physical, psychosocial, and cognitive) data. The aggregation of individual datasets in such a platform maximizes their utility and enables joint analyses of complex data, which increases power and provides a shared and secure environment without the risk of disclosing sensitive information.

Individual genetic data are not easy to share between studies due to the EU's General Data Protection Regulation (GDPR), in which genetic data are included in the list of sensitive data. Only secure computational platforms (like DPUK) with a legally compliant (ISO 27001) process of data handling and processing offer an opportunity to combine the genetic data from a number of studies.

Access to individual levels of genetic data provides a new independent resource not only to explore neurodegenerative diseases such as different types of dementias, Parkinson's disease (PD), and amyotrophic lateral sclerosis (ALS) from different research angles, but also perform joint analyses with the aim to uncover additional genetic associations and/or insights into relevant biological mechanisms. Recent advances in genome wide association studies (GWAS) have made an enormous contribution and provide valuable insights about the pathogenesis of neurodegenerative disease, providing a positive step forward for the development of disease-modifying treatments.[2] A polygenic risk score (PRS) approach that combines small additive effects of specific loci across the genome has become an increasingly powerful tool to help identify individuals at higher/lower risk of developing complex disorders. Furthermore, a PRS approach could also help explain the proportion of genetic variance that seems to be missing when focusing only on genome-wide significant hits. It has shown great potential in Alzheimer's disease (AD) prediction with accuracy[3,4] can be used for studying genetic overlap among disorders of the brain.[5,6]

The DPUK platform is a unique collection of studies which were historically collected in the UK over the past 50 years to answer specific research questions. The studies are complementary to other large UK cohorts (UK Biobank,[7] Genomics England[8]). With a rapidly increasing number of GWAS studies, there is a lack of independent studies that can be used for replication, polygenic risk scoring, and other analyses requiring sample independence. Until recently, the DPUK platform has been a large, valuable, but underused resource. The lack of homogeneity of the phenotypic and genotypic information makes it difficult to use and therefore data harmonization is crucial to leverage its full potential.

In this paper, we set an example of combining genetic data across five studies that were approved for this project and provide research-ready datasets to the wider community that can be compared and/or analyzed together. This has been achieved by the creation and installation of standardized processing pipelines on the DPUK Portal including quality control (QC) steps, genetic imputation, and calculation of stan-


RESEARCH IN CONTEXT

1. **Systematic review**: The authors have undertaken a comprehensive review of the literature using traditional (e.g., PubMed) sources. The relevant references were added to the paper describing the DPUK portal, cohorts, and data analysis methodology.

2. **Interpretation**: We generated and installed pipelines within the DPUK portal for quality-control, genetic imputation, and polygenic risk score (PRS) calculation. Pipelines, imputed genetic data, and PRS will be available for investigators via the DPUK platform, where individual study data access consent and pre-approved ethics permit such data sharing (upon data owner approval).

3. **Future directions**: Given the important value of data sharing from both a scientific and funder's perspective, it would be inappropriate for the scientific community not to continue offering and using these valuable resources, while ensuring compliance with the permissions and ethics of individual studies. This work allows the wider research community to access and process data at scale and pace.

---

dardized PRSs with the six latest GWAS summary statistics related to neurodegeneration diseases, namely AD,[9] AD-by-proxy, PD,[10] frontotemporal dementia,[11] ALS,[12] and Lewy body dementia.[13] All generated and QC-ed data are provided in a widely accepted PLINK format.[14] The pipelines are set as a series of commands in a bash script and can be easily modified if any additional data filtering is required. To perform other genetic analyses, software packages can be requested to be installed by the DPUK technical support team. Detailed information about the data application process to access DPUK cohorts and processed data is available on the DPUK portal (https://portal.dementiasplatform.uk/Apply). The associated phenotypic data processing and harmonization is ongoing. The ready-to-use, harmonized, and QC-ed data offers an advantage to researchers to accelerate collaborative projects, remove the need to repeatedly curate the data on per-project basis, and reduce the cost of data management and the level of uncertainty in the choice of analytical methodology. All data, pipelines, and PRS scores can be accessed and used within the DPUK platform by other researchers.

2 | METHODS

2.1 | Access data on DPUK

Bona fide academic and industry researchers are allowed to apply for access to the DPUK cohort datasets. Upon approval of an application and signing of a Data Access Agreement
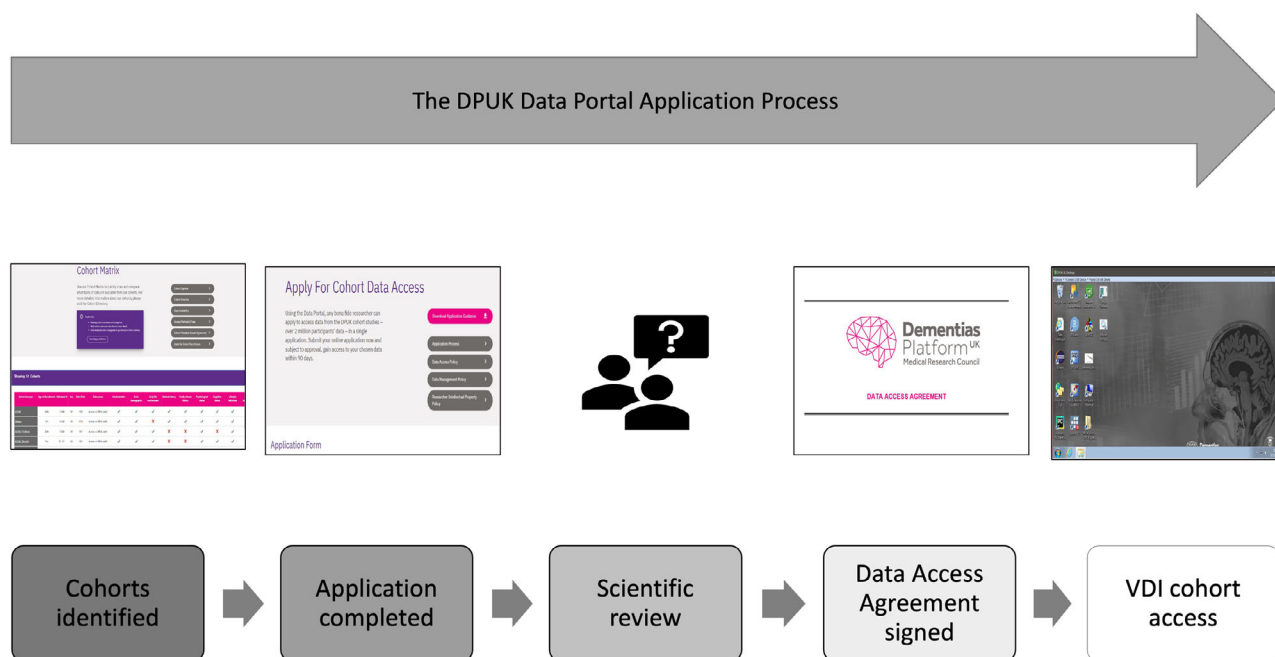
FIGURE 1 Flowchart of DPUK application process. DPUK, Dementias Platform UK; VDI, virtual desktop interface.

(https://portal.dementiasplatform.uk/Apply), researchers access approved datasets on a virtual desktop interface (VDI) within the DPUK Data Portal. All statistical packages and tools are preinstalled in the VDI and data cannot be downloaded. Figures, summary statistics, and graphs may be downloaded for publication and presentation purposes. Scripts may be uploaded onto the VDI. The flowchart of DPUK application process can be seen in Figure 1.

2.2 | Studies with genetics

For this project we used DPUK cohorts that agreed to participate in sharing the individual-level genotype data with the main aim of merging and processing the datasets together. These cohorts were also used to test the data processing pipelines and provide research-ready datasets for analyses by the individual cohorts, thereby encouraging collaboration among the studies. All cohorts had basic demographic information (sex, age, ethnicity), and most of the cohorts had cognitive tests and neurodegenerative disease diagnoses (clinical or *post mortem*). The cognitive assessments, however, were measured using different questionnaires, depending on the purpose of the study. The work to harmonize and standardize the phenotypic cognitive information is ongoing.

Ethical approval was not required as this was obtained at source by the cohort and only secondary analysis was undertaken.

Brains for Dementia Research (BDR)[15,16] is an initiative that has recruited participants across five UK brain banks to help to investigate the mechanistic pathways of dementia by studying phenotypic data collected during their lives and their donated brain tissue after death. BDR data collection is ongoing with > 3200 people signed up to donate their brains. We used the BDR data freeze as of October 2020,

including participants aged 56 to 104. The data collection has followed standardized operating procedures of brain donations along with standard longitudinal clinical and psychometric assessments and genetic data.

Generation Scotland (GS) of the Scottish Family Health Study is a family-based genetic epidemiology study with DNA and socio-demographic and clinical data from > 20,000 volunteers across Scotland aged 18 to 98 years, from February 2006 to March 2011.[17] Participants and their families were invited to take part in the study with the aim to investigate links between genetics and common complex familial diseases such as cardiovascular disease, cognitive decline, mental illnesses, and so forth.

Epic Norfolk (EN) is a part of the European Prospective Investigation into Cancer (EPIC), a large multi-center cohort study with participants enrolled from 23 centers across Europe, EN being one of them. More than 30,000 people living in Norwich and surrounding towns and rural areas were recruited into the EN study between 1993 and 1997 who were aged between 39 and 79.[18] The data include dietary and lifestyle information, health questionnaires, numerous disease diagnoses, and genetics.

The Medical Research Council National Survey of Health and Development (NSHD) is the longest-running British birth cohort (1946) from England, Scotland, and Wales. Five thousand three hundred sixty-two participants were recruited at birth in a single week in March 1946,[19] with > 2800 people in the active sample. Information that has been collected includes lifestyle, environmental, childhood health and development, lifetime social circumstances, genetic, and imaging data.

The Airwave Health Monitoring Study (AW) is a longitudinal epidemiological study of the police force to evaluate possible health risks associated with use of TETRA, a digital communication system used

TABLE 1 Genetic description of cohorts.

Cohort abbreviation	Cohort full name	Genetic data received	N SNPs	N samples	Phenotypes
BDR	Brains for Dementia Research	Neurochip	478,633	570	Sociodemographic, cognitive status, mental health, MMSE
GS	Generation Scotland	GS_SFHS, CHR(X), *APOE*	604,858 17,574,2	20,032 20,110	Sociodemographic, mental health, cognitive test, cognitive status
EN	EPIC Norfolk	Axiom 2020	728,244	21,041	
NSHD	Medica Research Council National Survey of Health and Development	NeuroX2 Imputed with HRC	11,081,207	2864	Amyloid status, brain measurements, MMSE
AW1	The Airwave Health Monitoring Study	Affymetrix	845,487	4493	Sociodemographic, cognitive status, mental health
AW2		IlluminacoreExome	542,677	14,887	

Abbreviations: *APOE*, apolipoprotein E; MMSE, Mini-Mental State Examination; SNP, single nucleotide polymorphism.

by police forces and other emergency services in Great Britain since 2001,[19,20] with 42,112 participants recruited by the end of 2012. The cohort has been richly phenotyped and has blood and urine samples, lifestyle factors, health screening, mental health, and well-being measurements and genetics. Summary of available genetics for these cohorts can be seen in Table 1.

2.3 | Genetic data harmonization

Before any joint genetic analysis, the data should be merged on overlapping single nucleotide polymorphisms (SNPs), harmonized, and checked for outliers. Originally, there were a total of 32,365 overlapping SNPs among five datasets (BDR, GS, EN, AW1, AW2) that were genotyped on different platforms (see Table S1 in supporting information). This significantly limits the capacity to conduct any genome-wide study at a SNP, gene, or haplotype level or construct PRSs across all studies.

We developed and installed a genotype QC and imputation pipeline to facilitate standardized procedures for all aspects of genetic data and it is now available on the DPUK platform. We have chosen a standard protocol[21] for QC analysis with widely used PLINK[14] and R software. The choice of thresholds for each QC step was not too stringent to retain the majority of individuals and genetic variants. However, (1) these thresholds can be adjusted within the pipeline if more stringent/relaxed inclusion criteria are required; (2) additional filtering steps can be applied by researchers on already QC-ed cohorts; and (3) additional software can be requested to be installed and applied to perform other genetic analyses, for example, to re-calculate kinship scoring.

The pipeline is initiated with pre-imputation QC checks that were applied to the all-target cohorts. Samples were removed based on call rate <95%; heterozygosity (HET > ± 0.1); relatedness based on identity by descent with PI_HAT > 0.2, except the GS cohort. We did not exclude related individuals in the GS sample, as the family members were specifically recruited according to the study design. All cohorts were merged with the 1000 Genomes dataset to conduct a principal component analysis (PCA). Individuals were removed if they did not cluster near the 1000 Genomes European cluster. SNPs were removed with minor allele frequency (MAF) < 0.01; Hardy–Weinberg equilibrium (HWE) $P_{HWE} \leq 10^{-6}$; with missing data proportion >5%. At the pre-imputation step, SNPs were aligned with the 1000 Genomes reference panel, hg19. SNP alignment included removing SNPs that have discordant information present with the reference panel (i.e., allele mismatch, strand flips, etc.). The pre-imputation QC steps and exclusions for each cohort are presented in Tables S2–S7 in supporting information and PCA are presented in Figure S1A–F in supporting information.
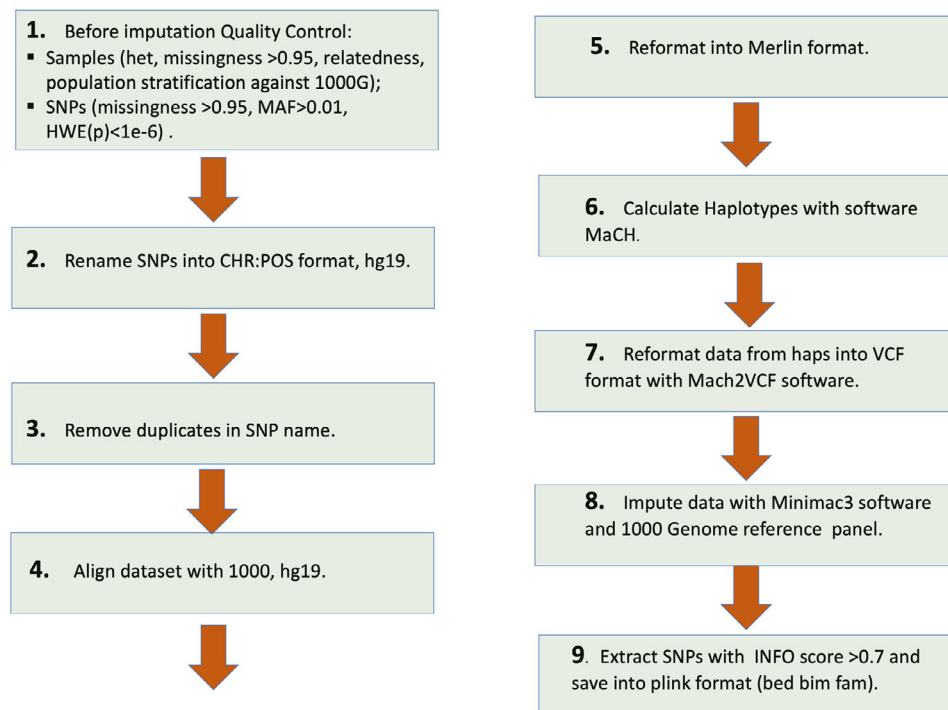
In the next step, the Minimac imputation tool[22,23] was implemented. This tool relies on a two-step approach: (1) phasing samples into a series of estimated haplotypes with MaCH software[24] and (2) using the derived haplotypes for genotype imputation. The 1000 Genomes reference panel (https://www.internationalgenome.org) in VCF format was used because it is publicly available for download onto the DPUK platform. We did not use HRC[25] or TOPMED[26] reference panels due to limitations induced by the data-sharing policy. The detailed workflow of the imputation protocol is represented in Figure 2.

The last step of the pipeline, post-imputation QC, was applied to remove variants with imputation information scores < 0.7, MAF < 0.01, and $P_{HWE} \leq 10^{-6}$.

2.4 | Derivation of PRSs

PRS derivation requires discovery GWAS summary statistics (effect sizes, reference alleles, and P values) and target data, which is independent of the GWAS with individual level genetic information available for each sample.

Before proceeding with PRS calculations, we uploaded to the DPUK Portal publicly available GWAS summary statistics for the six largest neurodegenerative disease studies: (1) clinical AD GWAS of 63,926 samples[9] (AD); (2) AD-by-proxy/clinical GWAS and related dementias

1. **Before imputation Quality Control:**
   - Samples (het, missingness >0.95, relatedness, population stratification against 1000G);
   - SNPs (missingness >0.95, MAF>0.01, HWE(p)<1e-6) .

↓

2. **Rename SNPs into CHR:POS format, hg19.**

↓

3. **Remove duplicates in SNP name.**

↓

4. **Align dataset with 1000, hg19.**

↓

5. **Reformat into Merlin format.**

↓

6. **Calculate Haplotypes with software MaCH.**

↓

7. **Reformat data from haps into VCF format with Mach2VCF software.**

↓

8. **Impute data with Minimac3 software and 1000 Genome reference panel.**

↓

9. **Extract SNPs with INFO score >0.7 and save into plink format (bed bim fam).**

FIGURE 2 Workflow of the imputation protocol for genotyped data. HWE, Hardy–Weinberg equilibrium; MAF, minor allele frequency; SNP, single nucleotide polymorphism.

(ADRD) of 487,511 samples;[27] (3) Parkinson's Disease GWAS (PD) of 1,474,097 samples;[10] (4) Frontotemporal Dementia GWAS (FTD) of 12,928 samples;[11] (5) Amyotrophic Lateral Sclerosis GWAS (ALS) of 138,086 samples;[12] and (6) Lewy Body Dementia GWAS (LBD) of 6618 samples.[13] In each set of GWAS summary statistics, we reformatted the variant IDs into "rs numbers," aligned them to the 1000 Genomes reference panel, and removed variants with standard error (SE) > 2 in the corresponding summary statistics. PRS was calculated for both all available SNPs and for all SNPs excluding *APOE* region (chromosome 19:44.4-46.5Mb) using AD and ADRD summary statistics (PRS.no.APOE).

Because there is still a debate about the comparability of various PRS approaches and optimal *P* value threshold, we have chosen the PRS approach with continuous shrinkage (PRS-CS)[28] that does not depend on *P* value threshold or clumping parameters and shows improved predictive accuracy across a wide range of disorders with complex genetic structure.[29] PRS-CS retains more SNPs and reduces information loss, compared to the widely used linkage disequilibrium (LD) clumping methods that only retain one lead SNP in an LD block.[30,31]

In the pipeline, PRS-CS scores were generated with six GWAS summary statistics for each cohort separately and on the combined dataset. The derived scores were adjusted for five principal components (PCs). We adopted the approach of PRS standardization, which allows scores to be comparable between studies.[31] For that, each cohort was merged with 1000 Genomes European population (N = 503) and we standardized the cohorts' PRS using the mean and standard deviation (SD) of the PRS from 1000 Genomes European population. The PRS calcu-

lation diagram can be seen in Figure 3. To investigate the difference between PRS distributions, the Kolmogorov–Smirnov test was applied, and *P* value was considered significant after Bonferroni correction for multiple testing ($P \leq 1.4e\text{-}3 = 0.05/36$).

3 | RESULTS

3.1 | Imputation

An overview of pre-imputation QC results, imputation, and post-imputation QC results that were performed for each cohort and final number of samples and variants are represented in Tables S2–S8 in supporting information. The six DPUK cohorts were imputed and QC-ed and are ready to be disseminated with pre-computed 5 PCs (with and without 1000 Genomes European population). The combination of six cohorts provides us with a dataset of 60,522 individuals on 4,037,483 variants, common among the cohorts.

3.2 | PRS for each study

Imputed and QC-ed genetic data was used for PRS score calculations and the scores are ready to be disseminated to other research projects. PRS-CS scores were generated for each cohort (BDR, GS, EN, NSHD, AW1, AW2), adjusted for PCs and standardized against 1000 Genomes European population, as described in Section 2. It can be observed that all PRS, as expected, have an approximately normal distribution; and
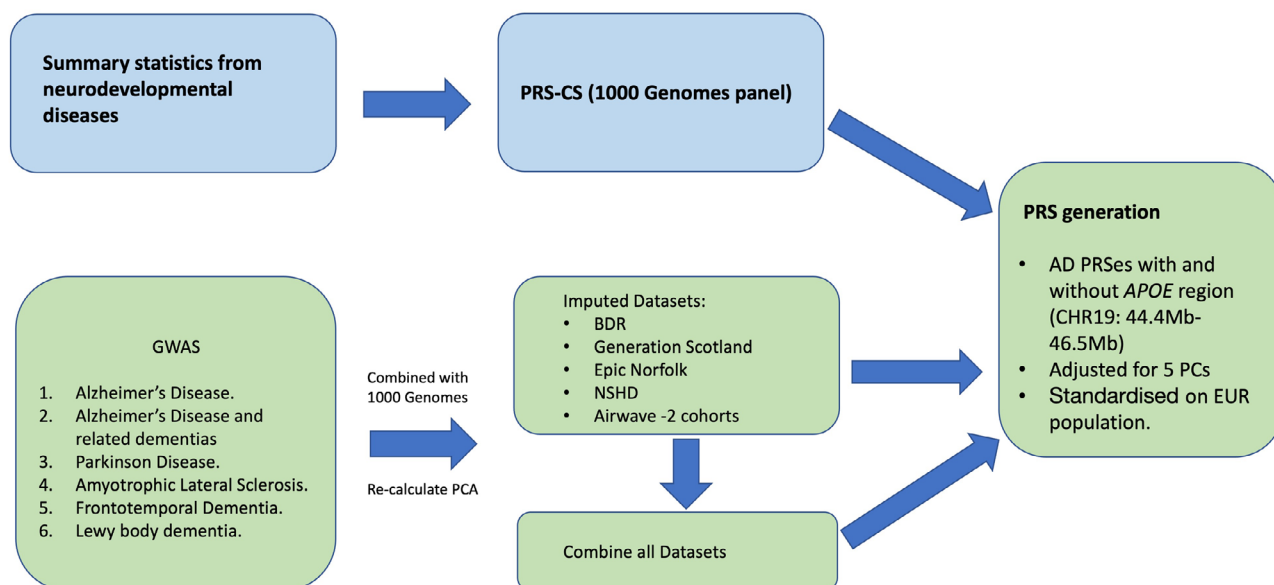
FIGURE 3 Workflow of the PRS generation protocol. *APOE*, apolipoprotein E; BDR, Brains for Dementia Research cohort; NSHD-MRC Medical Research Council National Survey of Health and Development cohort; PCA, principal component analysis; PC, principal components; PRS-CS, polygenic risk score approach calculated with 1000 Genomes reference panel.

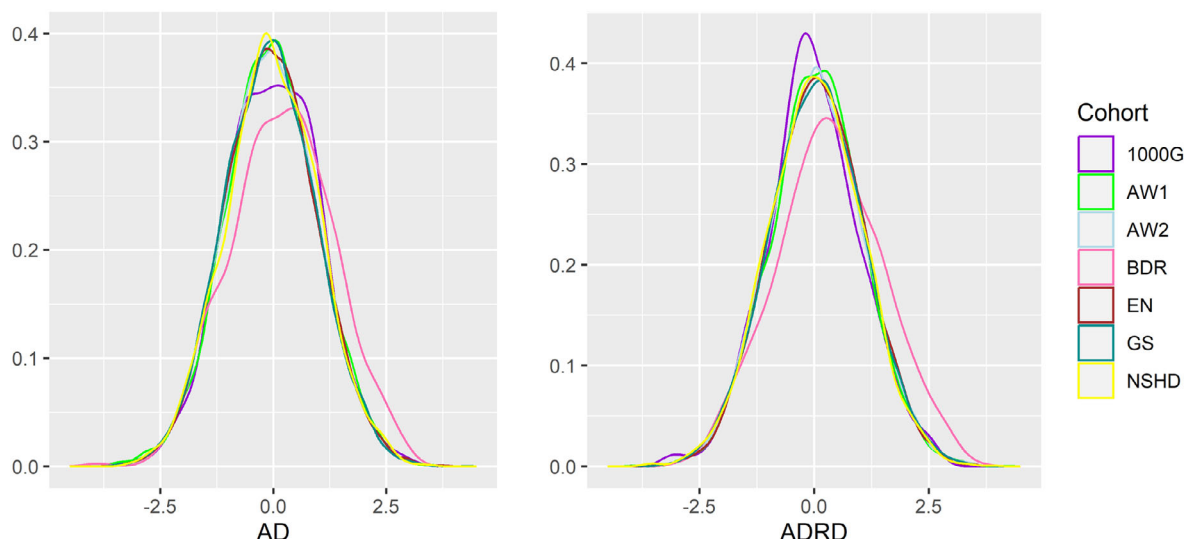cohorts' and European 1000 Genomes' PRS distributions are closely matched; see Figure S2A–F in supporting information.

3.3 | PRS distributions in combined study

First, we examined Pearson's correlations among all PRS-CS scores that were calculated for six neurodegenerative diseases. Figure 4 shows that the highest correlations (*r* between 0.34 and 0.91) can be observed between PRS calculated with AD and ADRD GWAS and depend on the inclusion of the *APOE* region. Correlation between AD and LBD PRS reached *r* = 0.11, while with other GWAS (PD, ALS) *r* is < ± 0.1. Note, that LBD-PRS correlates the most with both AD/ADRD and PD-PRS (0.11 and 0.09, respectively) and is in line with LBD diagnosis,[13] in which people with LBD have problems with understanding, thinking, memory, and judgement, similar to AD.

Next, we investigated PRS distributions of the combined dataset generated with six neurodegenerative GWAS (AD, ADRD, PD, ALS, FTD, LBD); see Figure S3A–F in supporting information with the corresponding Kolmogorov–Smirnov test *P* values in Table S9 in supporting information. Figure 5 presents standardized PRS distributions calculated with AD and ADRD summary statistics for each DPUK cohort. All PRS have similar to 1000 Genomes (purple line) normal distribution, with the exception of the BDR study (pink line) that is shifted to the right in both cases. Indeed, BDR is a case–control study (with pathologically confirmed diagnosis) and is enriched with dementia cases compared to other cohorts, which are population based. The difference between PRS distributions (BDR and 1000 Genomes) is border-line significant ($P = 6.5 \times 10^{-3}$) with AD-PRS and significant ($P = 1.1 \times 10^{-5}$) with ADRD-PRS; see Table S9.
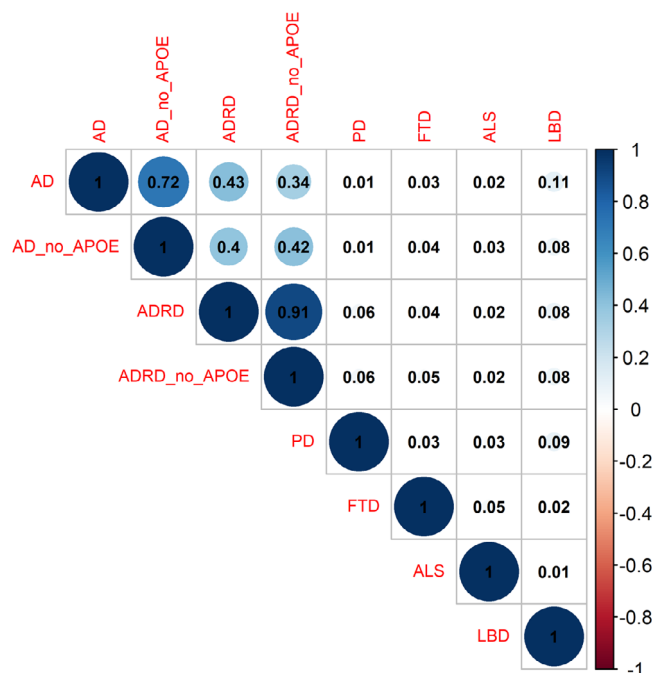
FIGURE 4 Matrix of Pearson's correlation of PRS-CS scores that have been calculated with six GWAS summary statistics: AD, ADRD, FTD, PD, ALS, LBD, with and without I region (AD, ADRD) in the combined cohort. *APOE*, apolipoprotein E; AD—clinical Alzheimer's disease GWAS; AD_no_APOE, Alzheimer's disease GWAS without *APOE* region; ADRD—Alzheimer's disease clinical/proxy GWAS and related dementias; ADRD_no_APOE—Alzheimer's disease clinical/proxy GWAS without *APOE* region; ALS, amyotrophic lateral sclerosis GWAS; FTD, frontotemporal dementia GWAS; GWAS, genome-wide association study; LBD, Lewy body dementia; PD, Parkinson's disease GWAS.

FIGURE 5 Standardized PRS distributions calculated with AD (left) and ADRD (right) summary statistics on combined dataset split by cohort (BDR, EN, GS, NSHD, AW1, AW2, 1000G). 1000G, 1000 Genomes European population cohort; AD, Alzheimer's disease; ADRD, Alzheimer's disease and related dementias; AW, Airwave Health Monitoring Study BDR, Brains for Dementia Research; EN, Epic Norfolk; GS, Generation Scotland; NSHD, Medical Research Council National Survey of Health and Development.

4 | DISCUSSION

The DPUK Data Portal has been designed to aggregate data from research groups across the United Kingdom and internationally into a single platform to maximize their utility and enable joint analysis of complex data that can lead to advancing new discoveries. Sharing genetic data is particularly challenging due to its identifiability, which requires protection and confidentiality but is of the utmost importance while requiring compliance with the permissions and ethics of each individual cohort. Given the complexity and heterogeneity of the genetic data due to genotyping platforms, differences in QC analyses, and the number of overlapping variants, when combined at the individual level, joint analysis is only possible after standardization and imputation of the data.

We have established a series of pipelines that involve (1) QC analysis prior to imputation, (2) imputation with the 1000 Genomes reference panel, (3) post-imputation QC analysis, and (4) calculation of PRS with the six latest and largest GWAS summary statistics of neurodegenerative disorders.

The data processing pipelines were installed with standard QC and data analysis parameters and are open-source scripts which can be easily adjusted by other researchers, suitable for the needs of their study designs. The pipelines can also be modified to perform other genetic analyses, that is, gene-set/pathway-specific PRS calculation with other GWAS summary statistics.

Our study has some limitations. First, for the PRS derivation, the independence between GWAS and the target dataset is required as even small sample overlap can produce significantly inflated results.[32] We were unable to analytically assess the sample overlap between GWAS and the DPUK datasets as only GWAS summary statistics are publicly available. However, to our best knowledge, there is no overlap between DPUK cohorts and the GWAS studies we have used.

Second, despite boosted statistical power, ADRD GWAS generated with clinically assessed AD cases that were meta-analyzed with "AD-by-proxy" approach[27] (AD diagnosis is based on participants' self-reported diagnosis for their parents) may have limitations that include imprecision of diagnosis, heterogeneity in the survey, and systematic biases related to UK Biobank sample collection.[33–35]

Third, the resulting number of SNPs shared between all DPUK cohorts is limited ($\approx$ 4 M), compared to other imputed datasets. This number is reduced because the NSHD study used NeuroX2 array for genotyping (with a small number of overlapping SNPs with any of the imputation reference panels). However, we provide imputed genetic data for each cohort separately on the DPUK Portal, which is equivalent to the expected number of imputed SNPs (8,9 million).

Finally, for the imputation, we have used the 1000 Genomes (publicly available) reference panel, as the DPUK data sharing policy does not allow any data to leave the platform, whereas the imputation with the TOPMED panel was only possible when the data moves to the Imputation Server provided by the University of Michigan (USA). We, however, used the same software and similar pipeline as implemented at the Michigan server.

In summary, imputed genetic data, the combined dataset, and PRS are now available for investigators via the DPUK Data Portal, where the individual study data access consent and pre-approved ethics permit such data sharing upon approved application. Given the important value of data sharing from both a scientific and funder's perspective, we encourage researchers to use these data as it would be inappropriate for the scientific community not to continue offering and using these valuable resources.

CONFLICT OF INTEREST STATEMENT

All authors have declared no conflicts of interest. Author disclosures are available in the supporting information.

CONSENT STATEMENT

All human subjects provided consent for participation with the source cohort. This consent included data collection and repurposing for secondary data analysis. Full ethical approvals had been obtained at source by the originating cohort according to their ethical approval body. For this study, additional ethical approval was not required as only secondary analysis was undertaken on anonymized secondary data from pre-consented human subjects.

ORCID

Ganna Leonenko https://orcid.org/0000-0001-8025-661X

Valentina Escott-Price https://orcid.org/0000-0003-1784-5483

REFERENCES

1. Bauermeister S, Orton C, Thompson S, et al. The Dementias Platform UK (DPUK) data portal. *Eur J Epidemiol*. 2020;35:601-611. doi:10.1007/s10654-020-00633-4

2. Chen W, Hu Y, Ju D. Gene therapy for neurodegenerative disorders: advances, insights and prospects. *Acta Pharm Sin B*. 2020:10:1347-1359. doi:10.1016/j.apsb.2020.01.015

3. Escott-Price V, Shoai M, Pither R, Williams J, Hardy J. Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. *Neurobiol Aging*. 2017;49:214.e7-214.e11. doi:10.1016/j.neurobiolaging.2016.07.018

4. Escott-Price V, Myers AJ, Huentelman M, Hardy J. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. *Ann Neurol*. 2017;82:311-314. doi:10.1002/ana.24999

5. Creese B, Vassos E, Bergh S, et al. Examining the association between genetic liability for schizophrenia and psychotic symptoms in Alzheimer's disease. *Transl Psychiatry*. 2019;9. doi:10.1038/S41398-019-0592-5

6. Bellou E, Stevenson-Hoare J, Escott-Price V. Polygenic risk and pleiotropy in neurodegenerative diseases. *Neurobiol Dis*. 2020;142. doi:10.1016/J.NBD.2020.104953

7. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3). doi:10.1371/journal.pmed.1001779

8. Caulfield M, Davies J, Dennys M, et al. National Genomic Research Library. figshare. *Dataset*. 2017. doi:10.6084/m9.figshare.4530893.v7

9. Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet*. 2019;51:44-430. doi:10.1038/s41588-019-0358-2

10. Nalls MA, Blauwendraat C, Vallerga CL, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol*. 2019;18:1091-1102. doi:10.1016/S1474-4422(19)30320-5

11. Ferrari R, Hernandez DG, Nalls MA, et al. Frontotemporal dementia and its subtypes: a genome-wide association study. *Lancet Neurol*. 2014;13:686-699. doi:10.1016/S1474-4422(14)70065-1

12. van Rheenen W, van der Spek RAA, Bakker MK, et al. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat Genet*. 2021;53:1636-1648. doi:10.1038/s41588-021-00973-1

13. Chia R, Sabir MS, Bandres-Ciga S, et al. Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture. *Nat Genet*. 2021;53:294-303. doi:10.1038/s41588-021-00785-3

14. Chang CC, Chow CC, Tellier LCCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4:7. doi:10.1186/s13742-015-0047-8

15. Francis PT, Costello H, Hayes GM. Brains for dementia research: evolution in a longitudinal brain donation cohort to maximize current and future value. *J Alzheimer's Dis*. 2018;66:1635-1644. doi:10.3233/JAD-180699

16. Young J, Gallagher E, Koska K, et al. Genome-wide association findings from the brains for dementia research cohort. *Neurobiol Aging*. 2021;107:159-167. doi:10.1016/J.NEUROBIOLAGING.2021.05.014

17. Smith BH, Campbell A, Linksted P, et al. Cohort profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int J Epidemiol*. 2013;42:689-700. doi:10.1093/IJE/DYS084

18. Hayat SA, Luben R, Keevil VL, et al. Cohort profile: a prospective cohort study of objective physical and cognitive capability and visual health in an ageing population of men and women in Norfolk (EPIC-Norfolk 3). *Int J Epidemiol*. 2014;43:1063-1072. doi:10.1093/IJE/DYT086

19. Wadsworth M, Kuh D, Richards M, Hardy R. Cohort profile: the 1946 National Birth Cohort (MRC National Survey of Health and Development). *Int J Epidemiol*. 2006;35:49-54. doi:10.1093/IJE/DYI201

20. Elliott P, Vergnaud AC, Singh D, Neasham D, Spear J, Heard A. The Airwave Health Monitoring Study of police officers and staff in Great Britain: rationale, design and methods. *Environ Res*. 2014;134:280-285. doi:10.1016/J.ENVRES.2014.07.025

21. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc*. 2010;5:1564-1573. doi:10.1038/nprot.2010.116.20105:9

22. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*. 2012;44:955-959. doi:10.1038/ng.2354

23. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48:1284-1287. doi:10.1038/ng.3656

24. Li Y, Willer CJ, Ding J, Scheet P, MaCH AbecasisGR. Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*. 2010;34:816-834. doi:10.1002/gepi.20533

25. The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48:1279-1283. doi:10.1038/ng.3643

26. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48:1284-1287. doi:10.1038/ng.3656.201648:10

27. Bellenguez C, Küçükali F, Jansen IE, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet*. 2022;54:412-436. doi:10.1038/s41588-022-01024-z

28. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun*. 2019;-1. doi:10.1038/s41467-019-09718-5

29. Ni G, Zeng J, Revez JA, et al. A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. *Biol Psychiatry*. 2021;90:611-620. doi:10.1016/j.biopsych.2021.04.018

30. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009;460:748-752. doi:10.1038/nature08185

31. Leonenko G, Baker E, Stevenson-Hoare J, et al. Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. *Nat Commun*. 2021;12. doi:10.1038/s41467-021-24082-z

32. Choi SW, Mak TSH, Hoggart CJ, O'reilly PF. EraSOR: a software tool to eliminate inflation caused by sample overlap in polygenic score analyses. *Gigascience*. 2022;12:1-11. doi:10.1093/GIGASCIENCE/GIAD043

33. Escott-Price V, Hardy J. Genome-wide association studies for Alzheimer's disease: bigger is not always better. *Brain Commun*. 2022;4:1-7. doi:10.1093/BRAINCOMMS/FCAC125

34. Grotzinger AD, la FuenteJde, Privé F, Nivard MG, Tucker-Drob EM. Pervasive downward bias in estimates of liability-scale heritability in genome-wide association study meta-analysis: a simple solution. *Biol Psychiatry*. 2023;93:29-36. doi:10.1016/J.BIOPSYCH.2022.05.029

35. Sun Z, Wu Y, Fetcher JM, Lu Q. Pervasive biases in proxy GWAS based on parental history of Alzheimer's disease. *Alzheimer's Dement*. 2023;19(Suppl.12):e080435. doi:10.1002/alz.080435

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Leonenko G, Bauermeister S, Ghanti D, et al. Dementias Platform UK: Bringing genetics into life. *Alzheimer's Dement*. 2024;20:3281–3289. https://doi.org/10.1002/alz.13782