The Polygenic Risk Score Knowledge Base: A Centralized Online Repository for Calculating and Contextualizing Polygenic Risk Scores

Madeline Page
University of Kentucky
Elizabeth Vance
University of Kentucky
Matthew Cloward
Brigham Young University
Ed Ringger
Brigham Young University
Louisa Dayton
Brigham Young University
ADNI Alzheimer's Disease Neuroimaging Initiative
Alzheimer&#x0092;s Disease Neuroimaging Initiative
Justin Miller
University of Kentucky
John Kauwe (✉ kauwe@byu.edu)
Brigham Young University

Article

Keywords: Polygenic risk score, GWAS Catalog, UK Biobank, 1000 Genomes, ADNI, genetic risk, genome-wide association study

Posted Date: September 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-799235/v1

Abstract

Introduction: Genome-wide association (GWA) studies identify correlation between genetic variants and phenotypes. GWA findings can be used to calculate polygenic risk scores, which represent the aggregate genetic risk across all associated loci.

Methods: We developed a centralized polygenic risk score calculator containing over 2,300 GWA studies from the NHGRI-EBI GWAS Catalog. Polygenic risk scores are calculated from user-uploaded data using various user-defined parameters across any disease(s) or studies.

Results: The Polygenic Risk Score Knowledge Base (https://prs.byu.edu) and command-line interface facilitate user-specific polygenic risk score calculations. We report study-specific polygenic risk scores across the U.K. Biobank, 1000 Genomes, and the Alzheimer's Disease Neuroimaging Initiative (ADNI) and identify potentially confounding genetic risk factors in ADNI.

Discussion: We introduce the first streamlined analysis tool and web interface to calculate and contextualize polygenic risk scores across various studies. We anticipate that the PRSKB will facilitate a wider adaptation and innovative use of polygenic risk scores in disease research.

Data Availability: This project is documented online at https://polyriskscore.readthedocs.io/en/latest/, and all programs are publicly available at https://github.com/kauwelab/PolyRiskScore. A web interface is also available at https://prs.byu.edu/.

Introduction

Genome-wide association (GWA) studies have revolutionized the study of complex diseases and trait heritability by identifying genome-wide significant genetic loci associated with specific phenotypes[1]. Over 71,000 genetic loci are currently implicated in diseases or traits with a genome-wide significant association ($p\text{-value} < 5x10^{-8}$)[2], and additional associations have been discovered through meta-analyses[3–5]. These studies span various complex diseases and traits, including major depressive disorder[6], type 2 diabetes mellitus[7], Alzheimer's disease[8], coronary artery disease[9], schizophrenia[10], numerous cancers[11–13], and lifestyle choices (e.g., smoking, drinking, etc.[14,15]). GWA studies have also helped determine the underlying genetic architecture of multiple complex diseases[16–18] and identified candidate drug targets[19–21].

However, a major limitation of GWA studies is that each genetic locus is evaluated individually. Polygenic risk scores address this issue by aggregating all effect size estimates implicated in a GWA study that are present in a genome of interest[22]. Complex diseases, such as Alzheimer's disease, often have dozens of associated genetic loci that each confer a relatively small protective or pathogenic effect, which a polygenic risk score would then sum to obtain an overall genetic risk for a specific genome.

Polygenic risk scores are dependent on the underlying summary statistics from a GWA study. However, most large-scale GWA studies have been conducted on predominantly European populations[23], with results that often do not translate to other populations due to differences in allele frequencies and linkage disequilibrium patterns[24–26]. For instance, a study of hypertrophic cardiomyopathy in African Americans misclassified benign genetic variants as pathogenic, which could have been avoided if even a small number of African Americans were included in the control group[27]. The lack of diversity in GWA study cohorts can also cause important risk alleles in minority populations to remain unidentified. For example, the Population Architecture using Genomics and Epidemiology (PAGE) study found that a novel risk variant associated with the number of cigarettes smoked per day existed at a frequency of 17.2% in Native Hawaiian participants but was absent or rare in most other populations[28].

Therefore, choosing an appropriate GWA study to calculate polygenic risk scores is paramount to the fidelity of the calculations because the accuracy and predictive power of a polygenic risk score is dependent on the power and scope of the corresponding GWA study data[29,30]. When used appropriately, polygenic risk scores can capture missing heritability and are a measure of genetic risk for a trait compared to the average risk in a specific population[31–34]. Because polygenic risk scores can stratify populations based on distinct risk, they are useful in determining clinical and personal interventions[35,36]. For example, a polygenic risk score can greatly inform cancer risk management for *BRCA1* carriers, who have a 21% risk of developing breast cancer by age 50 if they are in the lowest polygenic risk score decile for breast cancer, and a 39% risk of developing breast cancer by age 50 if they are in the highest polygenic risk score decile[37]. Likewise, polygenic risk scores are used to classify disease subtypes[36,38,39]. For instance, nominally significant differences in genetic risk scores for bipolar disorder between patients with and without psychotic symptoms suggest a more valid subclassification of the disease[40,41]. Furthermore, polygenic risk scores can effectively explore genetic overlap between pairs of traits[42], which has revealed a shared genetic basis for multiple pairs of psychiatric disorders[43,44] and a lack of correlation in pairs of neurological traits, such as multiple sclerosis and amyotrophic lateral sclerosis, where genetic correlation might otherwise have been expected[45]. Polygenic risk scores can also test for gene-by-environment and gene-by-gene interactions[46,47] through Mendelian randomization studies, which detect causal genetic relationships[48,49].

There currently exists a spectrum of tools available for calculating polygenic risk scores, ranging from direct-to-consumer genetic companies (e.g., 23andMe[50]) to downloadable software packages (e.g., PRSice-2[51]). PRSice-2 is a multi-faceted tool that greatly facilitates polygenic risk score analyses of large cohorts, but it requires users to have an in-depth knowledge of bioinformatics, supply their own GWA summary statistics, use bgen file format or binary PLINK[52] format for genetic data (i.e., no VCF files), and perform all calculations locally (i.e., no dedicated server for testing and/or small datasets). Additionally, PRSice-2 requires extensive startup time to fully utilize all available options, and it does not allow users to contextualize their results against pre-computed results from large independent biobanks[51]. These constraints have potentially limited the application of polygenic risk score calculations in assessing off-target disease susceptibility and the wider adaptation of polygenic risk scores in other genetic analyses.

Here, we present the Polygenic Risk Score Knowledge Base (PRSKB), a web server (https://prs.byu.edu) and command-line tool for calculating polygenic risk scores. PRSKB currently contains 2,388 previously published GWA studies spanning 862 diseases or traits. The PRSKB allows users to quickly compare polygenic risk scores calculated across diverse studies and contextualize those results against larger cohorts. We employ the 1000 Genomes database[53], UK Biobank[54], and Alzheimer's Disease Neuroimaging Initiative (adni.loni.usc.edu) to create polygenic risk score percentiles, against which individual reported risk scores can be examined.

We anticipate that the PRSKB will enable a wider adaptation of polygenic risk score calculations (e.g., clinical trial screening, analysis of comorbidities, identifying confounding genetic factors, analyses of common disease genetics, etc.) because it simplifies polygenic risk score calculations and contextualization across thousands of studies.

Methods

Data compilation

The PRSKB integrates with the National Human Genome Research Institute-European Bioinformatics Institute (NHGRI-EBI) GWAS Catalog[55] to provide the most up-to-date and comprehensive list of GWA studies. The GWAS Catalog is a publicly available database of GWA study summary statistics, which includes over 4,000 publications, 130,000 variant-trait associations, and 6,000 full summary statistic files. Study and association data from the GWAS Catalog are automatically downloaded, pruned, and reformatted using the gwasrapidd R library[56]. The data are filtered to include only associations that contain both an odds ratio and the respective risk allele. Additionally, only non-haplotype associations that reside on an autosomal chromosome are preserved. Finally, any allele that has been reported on the reverse strand is automatically detected and flipped to the forward strand. The strand-flipping procedure entails comparing each reported risk allele to the list of possible alleles for the specified variant from dbSNP[57]. If the reported risk allele does not exist in the list of possible alleles, the complement of the risk allele is checked against the list. If the complement is present, then it is used as the reported risk allele for polygenic risk score calculations, as recommended by Choi, et al. [22].

PRSKB tool structure

The PRSKB is divided into three key parts: the database, the server, and the client as shown in Figure 1. The GWA study data, linkage disequilibrium clumping data, and association data are housed in a MySQL database on the PRSKB server. Supplemental Tables S1-S3 expound on the information found in each database table. The variant associations from each study/trait combination are contained within a single *associations table*, which includes detailed summary statistics for each variant (see Supplemental Table

S1). The *study table* (see Supplemental Table S2) contains detailed descriptions of each GWA study. Finally, there are four clumps tables, *hg38 clumps*, *hg19 clumps*, *hg18 clumps*, and *hg17 clumps*, that include linkage disequilibrium region identification numbers for variants in each of the five super populations from the 1000 Genomes project (see Supplemental Table S3). The associations and study tables are automatically updated monthly with new associations added to the GWAS Catalog. The scripts for loading tables into the database are publicly available at https://github.com/kauwelab/PolyRiskScore/tree/master/update_database_scripts.

The server houses the application programming interface (API) endpoints for the PRSKB, running NodeJS using PM2 (https://pm2.keymetrics.io/) and NGINX (https://www.nginx.com/). While the user does not interact directly with the API endpoints, the client calls endpoints to download requested data needed to calculate polygenic risk scores. All calculations occur client-side to reduce strain on the server.

Users have two platforms from which they can calculate polygenic risk scores. The first platform is a web interface accessible at https://prs.byu.edu via a web browser. The second is a command-line interface (CLI) tool that can be run from the Linux or Mac command-line or from a bash shell on Windows. The CLI includes a bash script and four Python scripts. We recommend using the CLI to calculate polygenic risk scores for multi-sample VCF files.

Linkage disequilibrium clumping

Linkage disequilibrium is the nonrandom association of alleles at two or more loci[58]. Linkage disequilibrium generally affects loci that reside in close physical proximity, resulting in the joint inheritance of alleles at different loci within families and populations. Genetic variants that are in linkage disequilibrium will be similarly associated with traits in GWA studies. If they are not adequately assessed, they can confound a polygenic risk score analysis by overrepresenting the relative risk for a disease. For example, if three loci are in high linkage disequilibrium, only one locus should be included in calculating a polygenic risk score because the same risk signal is present in any of those three loci.

To reduce inflated polygenic risk scores, the genetic variants used to calculate polygenic risk scores need to be largely independent from each other. Linkage disequilibrium was calculated by first separating the 1000 Genomes data into the five previously annotated super populations: African, American, East Asian, European, and South Asian. We then used PLINK Linkage Disequilibrium (LD) Clumping[59] to calculate linkage disequilibrium regions for the variants in each population (see Supplemental File S1). We ran this analysis for the data available in both reference genomes hg38 and hg19. Although linkage disequilibrium regions are nearly identical between reference genomes[60], we also converted the variant coordinates in each clump to reference genomes hg18 and hg17 so that user-supplied genotype information could be easily mapped to the correct LD clump regardless of reference genome.

The LD Clumping analysis results were subsequently used to assign each genetic variant to an LD clump identifier (clump ID) for each population. LD regions were determined using an r-squared cutoff of 0.25 and a distance threshold of 500 kb, which correspond to parameters used in previous studies[61,62]. From this information, we created a table of population-specific linkage disequilibrium clusters for each reference genome in our database (see Supplemental Table S3). The clump ID for each population facilitates the dynamic retrieval of LD clumps from the database so that no more than one variant per LD region is included in a polygenic risk score calculation.

Calculating Polygenic Risk Scores

Polygenic risk scores are calculated using the same protocols outlined by Choi, et al. [22]. Figure 2 shows that polygenic risk score calculations require two essential datasets: 1. summary data comprised of GWA study summary statistics (e.g., odds ratios and p-values), and 2. user-supplied query data comprised of individual genotypes. Although a single GWA study is used to calculate each polygenic risk score, users can optionally select multiple studies or traits, which will each be analyzed independently. Users can also upload their own GWA summary statistics for personalized analyses. The PRSKB first ensures that the summary data and the query data are in the same format (e.g., strand flipping and same reference genome). Next, linkage disequilibrium is calculated by comparing each locus to the population-specific clumping regions housed on the server. When an individual in the target data has two or more variants within the same clumping region, the PRSKB chooses the variant with the most significant p-value from that region to represent the clump in the polygenic risk score. After adjusting for linkage disequilibrium, the remaining set of variants that overlap between the sample and the GWA study are retained and used in the polygenic risk score calculation. The equation we use in the PRSKB to calculate genetic risk scores sums the weighted effects of the overlapping alleles present in an individual, as seen in Equation 1. The value for a_i is the number of risk alleles at the i^{th} locus, r_i is the odds ratio at that locus, and b and c are the coefficient and exponent for the p-value threshold, respectively. The p-value threshold is supplied by the user at runtime.

Equation 1:

$$PRS = exp\left(\sum_{0}^{i}\begin{cases} a_i * \ln(r_i)\,, p < b * 10^c \\ 0, p \ \geq b * 10^c \end{cases}\right)$$

We chose to implement this equation because scores calculated in this simple manner are generally highly accurate[9,22,26,29,63,64]. Although the additive polygenic risk score model does not account for gene-gene or gene-environment interactions, the largest meta-analysis of heritability from twin studies validates a simple additive model for a majority of the traits examined[65].

UK Biobank and 1000 Genomes Polygenic Risk Score Visualization

In order to adequately interpret polygenic risk scores, individual results must be contextualized against a large dataset of similar ethnicity[29]. We used the 1000 Genomes Project[53], which contains sequencing data for 2,504 samples spanning five different continental regions, and the UK Biobank[54], a biomedical database comprised of genetic data for 500,000 individuals from the United Kingdom, to generate risk score distributions and summary statistics for each study in the PRSKB database. First, we divided the samples by population region. Next, we used the PRSKB to compute polygenic risk scores from all GWA studies in our database for each individual in each dataset. We then calculated the percentile rank of each person against all other people in the dataset with the same ethnicity. The polygenic risk score and percentile ranks were passed to Plotly JavaScript[66] to create interactive graphics that allow users to visualize population-specific distributions of polygenic risk scores for any study in the PRSKB database. Dynamic plots with a table of summary statistics for each study are available for users to query online at https://prs.byu.edu/visualize.html.

Alzheimer's Disease Neuroimaging Initiative (ADNI) Case Study

We also computed Alzheimer's disease polygenic risk scores for individuals in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) to verify the efficacy of the PRSKB calculations. 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. MCI is the preclinical stage of Alzheimer's disease, and is characterized by a slight but measurable decline in cognitive abilities. Individuals with MCI are at an increased risk of developing Alzheimer's disease or another dementia.

The ADNI case-control cohort includes clinical and genetic data for over 5,000 individuals. Results of a multiple linear regression indicated that there is not a collective significant relationship between age, sex, and polygenic risk score (F=0.395, p=0.674) in the ADNI dataset. Further examination revealed that neither age (t=0.0256, p=0.666) nor sex (t=0.614, p=0.469) has significant individual association with polygenic risk score outcomes, so we did not adjust for age or sex in our analyses using the ADNI dataset. We used the PRSKB calculator to compute the polygenic risk scores for each individual, and we chose to use the most comprehensive GWA study on Alzheimer's disease in the PRSKB database[4] as the summary data for calculations.

Following the risk score calculations, we separated the individuals by clinical dementia rating (CDR)—a summary measure developed to denote the overall severity of dementia in an individual, where CDR=0.0 is cognitively normal, CDR=0.5 is MCI, and CDR≥1.0 is Alzheimer's disease[67]. A Shapiro-Wilk's test of normality[68] revealed that the risk scores in each group were not normally distributed (Alzheimer's disease $p=2.2\times10^{-16}$, MCI $p=1.4\times10^{-7}$, cognitively normal $p=4.5\times10^{-10}$), so we opted to use a Mann-Whitney U test[69] to compare the distributions of polygenic risk score ranks in individuals with and without Alzheimer's disease. We first compared genetic risk scores in individuals with a CDR≥1 (Alzheimer's disease) and individuals with a CDR≤0.5 (MCI + cognitively normal). Next, we compared individuals with a CDR=0 (cognitively normal) and individuals with a CDR≥0.5 (Alzheimer's disease + MCI). Because the shape of the risk score distributions in each group followed a similar pattern, the Mann-Whitney U analyses were able to determine the extent to which the medians in each group significantly differed.

We performed a similar analysis using every study in the PRSKB database to identify additional traits that have distinct polygenic risk score distributions for each CDR cohort in the ADNI dataset. Using the PRSKB, we generated polygenic risk scores from 2,388 studies for each individual in the ADNI cohort. We then separated the individuals by CDR diagnosis, where individuals with a CDR≥1 (Alzheimer's disease) were placed in one group and individuals with a CDR≤0.5 (MCI + cognitively normal) were allocated to another group. Next, we used a Shapiro-Wilks[68] test to determine the extent to which scores for each study/CDR cohort were normally distributed. For normally distributed data, we performed an independent Welch's two-sample t-test to compare the risk score means of individuals with and without Alzheimer's disease. Finally, in studies with non-normal distributions of polygenic risk scores, we performed the non-parametric Mann-Whitney U test[69], which compared the distribution of ranks in individuals with and without Alzheimer's disease. Cohorts with a sample size less than two were excluded from the analysis. We repeated this analysis using a different clustering: CDR≥0.5 (Alzheimer's disease + MCI) in one group and individuals with a CDR=0 (cognitively normal) in another group. We did not analyze MCI as a separate group to maintain statistical power and identify differences between 1) People with Alzheimer's disease and all other individuals and 2) People with normal cognition and all other individuals. Similar to the computations performed with the UK Biobank and 1000 Genomes datasets, we also report the percentile score distributions and summary statistics for CDR≥1, CDR=0.5, and CDR=0 online using Plotly Javascript[66].

Results

We developed the PRSKB to simplify the process of calculating polygenic risk scores across all available GWA studies. Users can calculate polygenic risk scores through the user-friendly online calculator or command-line interface. The PRSKB GWA Study Browser allows users to choose which GWA studies to use for calculations and provides references for each study. Polygenic risk scores can be contextualized against the UK Biobank, population-specific 1000 Genomes data, and ADNI dataset for each study in the database. The depth and breadth of studies in the database, as well as the collection of previously calculated risk scores from a variety of populations, facilitates the implementation of the PRSKB in future trait and disease research.

Polygenic Risk Score Calculator Online

The PRSKB calculator can calculate polygenic risk scores for multiple traits and studies. To run the calculator, users input target data either by typing reference SNP ID numbers and their corresponding alleles into a text box or by uploading a variant call format (VCF) file. Next, the user must specify the reference genome (hg38, hg19, hg18, or hg17) used to sequence the input variants so that the associations queried from the database correspond to the same reference assembly. The user must also indicate the population of the samples to perform accurate linkage disequilibrium clumping. Various filters allow the user to choose specific studies, ethnicities, or study types (e.g., the highest impact determined by Altmetric score[70] or the largest study cohort measured by the initial sample size plus the replication sample size). Finally, the user must designate a p-value threshold for GWA variants included in the calculations and whether they prefer a condensed or verbose output file. Supplemental Figure S2 presents the PRSKB calculator interface.

The polygenic risk score results are written to a tab-separated values (TSV) output file presented in either a condensed or detailed format, or a JavaScript Object Notation (JSON) file (see Supplemental Figure S3). Supplemental Figure S4 displays an example of a condensed TSV output file, where each row includes the GWAS Catalog accessions ID, trait, trait description, GWA study citation, and the corresponding polygenic risk score for each sample. The verbose TSV output format is shown in Supplemental Figure S5, where each row contains the sample name, GWAS Catalog accession ID, GWA study citation, trait, trait description, corresponding polygenic risk score, and lists of the variants that decrease risk, increase risk, and have neutral or unknown effect on the polygenic risk score. Genetic variants with an odds ratio greater than one indicate an increased risk of developing the disease or trait, while odds ratios less than one indicate genetic protection against the disease or trait.

Users can browse the GWA studies in our database to locate studies they wish to use in their calculations by searching for the first author, article title, trait, PubMed ID, or GWAS Catalog study accession ID. The GWA study browser can be accessed under the "Studies" tab on the PRSKB website or through "Option 2: Search for a specific study or trait" on the PRSKB CLI menu. Supplemental Figure S6 introduces the GWA study browser interface.

Command-line Interface Tool Download

In addition to the website, a downloadable command-line interface (CLI) tool is available for users to run the calculator directly from the command-line. Required parameters include a path to the input file, a path to the output file, the p-value threshold for associations, the reference genome of the variants in the input file, and the super population for the samples in the input file. Using only the required parameters, genetic risk score calculations are run on every trait and study in the database. Optional parameters are used to filter which studies are included for calculations (e.g., specific traits, studies, or ethnicity of the study cohort). The CLI can also be run in two steps to perform large calculations without internet access, and it is multithreaded for improved computational efficiency (see Supplemental Figure S7).

The CLI tool contains a built-in menu when run without parameters. This menu allows users to learn more about the CLI tool and the parameters required to run it, search the PRSKB database for traits and studies, view the usage statement, and run the risk score calculator (see Supplemental Figure S8).

The UK Biobank, 1000 Genomes, and ADNI for Genetic Risk Score Contextualization

We present polygenic risk score distributions and summary statistics for each of the studies in the PRSKB database, generated from individual genetic data in the 1000 Genomes, UK Biobank, and ADNI datasets.

Users can choose between the following cohorts as an approximate contextualization for their own reported risk scores: UK Biobank, 1000 Genomes—African, 1000 Genomes—American, 1000 Genomes—East Asian, 1000 Genomes—European, 1000 Genomes—South Asian, ADNI—Alzheimer’s disease, ADNI—Mild Cognitive Impairment, and ADNI—cognitively normal. Polygenic risk score distributions can be visualized as violin plots, box plots, or line plots of the percentile data. For example, Supplemental Figure S9 depicts the distribution of SARS-CoV-2 polygenic risk scores for individuals in the UK Biobank cohort based on GWA summary statistics reported by Ellinghaus, et al. [71].

ADNI Case Study

Although we compared only two groups in the ADNI dataset (i.e., Alzheimer's disease or mild cognitive impairment versus controls and controls or mild cognitive impairment versus Alzheimer's disease) because of limited sample size for mild cognitive impairment, we used an adjusted significance level of 0.01 to account for multiple testing of five potential comparisons of Alzheimer's disease risk (i.e., Alzheimer's disease versus mild cognitive impairment; Alzheimer's disease versus controls; mild cognitive impairment versus controls; Alzheimer's disease or mild cognitive impairment versus controls; and mild cognitive impairment or controls versus Alzheimer's disease). A Mann-Whitney U test revealed a significant difference between Alzheimer’s disease polygenic risk score medians in individuals with a CDR≥1 and individuals with a CDR≤0.5 (U=67845; p=0.00332). Conversely, a Mann-Whitney U test did not detect a significant difference between Alzheimer’s disease polygenic risk score medians for individuals with a CDR=0 and individuals with a CDR≥0.5 (U=42191; p=0.04273). Figures 3 and 4 show the comparisons of polygenic risk score distributions in each CDR cohort.

After calculating polygenic risk scores from all other studies in the PRSKB database for the individuals in the ADNI cohort and correcting for multiple testing, we detected 17 GWA studies that had polygenic risk score means or medians that were significantly different ($p<2.36x10^{-05}$) between individuals with and without Alzheimer’s disease (see Supplemental Table S4). For example, a Mann Whitney U test revealed a significant difference (U=82626; $p=4.57x10^{-05}$) between polygenic risk score medians for ischemic stroke (large artery atherosclerosis)[72] in individuals with a CDR≥1 (Alzheimer’s disease) and individuals with a CDR≤0.5 (i.e., MCI + cognitively normal). However, a Mann Whitney U test did not indicate a significant difference (U=50882; $p=1.27x10^{-3}$) between polygenic risk score medians for ischemic stroke[72] in individuals with a CDR≥0.5 (i.e., Alzheimer’s disease + MCI) and individuals with a CDR=0 (cognitively normal) (see Supplemental Table S5).

Discussion

The PRSKB is the bridge between GWA study data and calculating polygenic risk scores using user-specific datasets. Polygenic risk score calculations require GWA study summary statistics, yet current tools for calculating polygenic risk scores do not offer straightforward, comprehensive access to usable GWA study information. The PRSKB facilitates large-scale polygenic risk score analyses currently spanning 2,388 GWA studies and 862 diseases or traits, which will likely enable researchers to identify previously unknown genetic biases in sampled cohorts and/or potential associations between traits.

The PRSKB also improves polygenic risk score utilization by offering contextualization for individual risk scores. The UK Biobank, 1000 Genomes, and ADNI genetic risk score percentiles provide the information necessary for users to normalize their reported scores relative to a large population-specific dataset.

The application of polygenic risk scores has become a critical resource in researching complex genetic diseases and personalized medicine. Although polygenic risk scores are effective at predicting genetic liability to a trait[31–34], risk prediction is not always the end objective to performing polygenic risk score calculations. Rather, these analyses are used for a wide variety of research purposes. Polygenic risk scores are useful at stratifying populations[35], influencing clinical and personal disease interventions[36,37], classifying disease subtypes[38,39], identifying genetic overlap between traits[42,44], and determining causal genetic relationships through Mendelian randomization studies[48,49,73]. Moreover, the implementation of polygenic risk scores has the potential to limit unknown covariates in future genetic studies by revealing individuals that have atypical genetic risk for phenotypes not directly studied.

Although polygenic risk scores have become increasingly prevalent in genetic research, historically, only minimal guidelines have existed for performing polygenic risk score analyses[22]. This limitation has led to inconsistencies in polygenic risk score methodologies in different studies and the misinterpretation of results. A recent publication by Choi, et al. [22] outlines a protocol for calculating polygenic risk scores, including detailed guidelines for performing and interpreting genetic risk score analyses. In our efforts to overcome the variability in current polygenic risk score research, we follow the standards set forth by Choi, et al. [22], including the implementation of the clumping and threshold (C + T) method. Furthermore, users are encouraged to follow the quality control measures for target and GWA data recommended by Choi, et al. [22] in order to ensure more optimal polygenic risk scores. Specifically, users are encouraged to ensure that the summary data and target samples are from the same population but avoid sample overlap or highly related samples. A target sample size of at least 100 and GWA study data with a SNP heritability (h^2_{SNP}) > 0.05 will also improve the power and accuracy of genetic risk score results[22]. Furthermore, we suggest that users who utilize the PRSKB to run bulk polygenic risk score analyses for post-hoc hypothesizing account for multiple testing when determining a significance threshold.

There are certain limitations to the PRSKB. First, while the PRSKB contains over 17,900 GWA variants, haplotype associations, or associations that include multiple variants for a single effect size, were removed from the PRSKB database. Currently, less than two percent of studies in the GWAS Catalog contain haplotype data. Additionally, certain studies in the GWAS Catalog have duplicate variants with varying p-value annotations. For example, the same variant in a study could have two different p-values: one labeled as 'male' and the other labeled as 'female.' In this case, the variant is excluded from calculations by default unless the user specifically opts to include gender specific variants. Other studies have variants with multiple p-values based on population description or other study specific designations.

Due to the complexity and lack of uniformity for these variant annotations, we have chosen to exclude those variants from calculations. Finally, although LD Clumping is the preferred method for the removal of variants in linkage disequilibrium[22], a common criticism of clumping is that the correlation and distance thresholds are generally arbitrarily chosen[22,61]. We selected threshold values that emulate clumping procedures performed in previous studies[61,62], but recognize that this choice may be an area for further development and research.

Additionally, the PRSKB tool has other limitations that are inherent in GWA studies and polygenic risk score calculations[74]. A common limitation of GWA studies is their current inability to account for more than a small fraction of complex trait heritability[75]. Much of this missing heritability is attributed to rare variants or variants with small effect sizes that do not reach genome-wide statistical significance[76]. Additional heritability has been uncovered over the last decade with the increase in GWA study sample size. For example, a 2009 study with 3,322 cases and 3,587 controls detected only a single genomic locus associated with schizophrenia[44], but by 2014, the number of genetic loci associated with schizophrenia had increased to 108 by using a sample size of over 36,000 cases and controls[77]. Although the number of variants identified have increased with GWA study sample size, the effect size for the majority of significant GWA loci is under 1.1, which makes it difficult to determine the individual functional effects of each identified variant[74]. A polygenic risk score confronts this matter by aggregating the individual effects of GWA study variants, but it also assumes that the genetic risk is additive.

The polygenic risk scores calculated for the individuals in the ADNI dataset reveal that the PRSKB is effective at estimating disease risk. As shown in Figs. 3 and 4, individuals with Alzheimer's disease displayed significantly higher genetic risk scores for Alzheimer's disease than individuals with MCI or who were cognitively normal. Recent findings by Leonenko, et al. [78] show that polygenic risk scores drive the severity of cognitive decline. Leonenko, et al. [78] demonstrated that the *APOE* gene was found to be the best predictor of amyloid deposition—a pathological hallmark of Alzheimer's disease and an important factor in neural degeneration. However, they also found that progression from amyloid accumulation and MCI to Alzheimer's disease was better determined by polygenic risk score, not *APOE* status. Our polygenic risk score calculations similarly show that polygenic risk scores are effective at capturing the distinction between MCI and Alzheimer's disease in the ADNI cohort.

The analyses on the ADNI cohort also highlight the utility of polygenic risk scores in identifying groups of individuals with distinct genetic risk for a certain trait. For example, a Mann Whitney U test revealed that genetic risk for ischemic stroke is significantly different between individuals with and without Alzheimer's disease ($U = 82626$; $p = 4.57 \times 10^{-05}$), as shown in Supplemental Table S4. Ischemic stroke is a known risk factor for Alzheimer's disease[79], and the two diseases share numerous pathophysiological mechanisms, particularly those influenced by inflammation, immune exhaustion, and neurovascular unit compromise.[80,81] Additionally, a study performed by Rahman, et al. [82] identified 22 unique genes that were dysregulated in both Alzheimer's disease and ischemic stroke datasets. Although the causal relationship between Alzheimer's disease and ischemic stroke is still unclear, our polygenic risk score analysis identified the shared genetic predisposition for these two traits in Alzheimer's disease patients from the ADNI dataset. Further examination is necessary to determine the extent to which genetic predisposition for ischemic stroke relates to Alzheimer's disease or if genetic predisposition for ischemic stroke is an independent covariate in the ADNI dataset. We anticipate that the PRKSB will allow polygenic risk score analyses of this kind to become a regular part of genetic cohort selection and downstream genetic analyses.

As GWA studies continue to improve, the polygenic risk score calculations computed in the PRSKB will become more powerful and effective. Recent efforts to recognize and improve the lack of diversity in GWA study sample populations[25,83] will allow users to compute polygenic risk scores for a wider range of ethnicities and help reduce population biases in polygenic risk score calculations. Furthermore, as GWA study sample sizes increase, additional loci with genome-wide association will be revealed, resulting in more comprehensive polygenic risk scores. Empirical evidence indicates that for each complex phenotype, there is a threshold sample size above which the rate of variant discovery increases dramatically[84]. Moreover, the detection of risk variants has yet to plateau for any trait[84], suggesting that as large cohorts become increasingly available, polygenic risk scores will become more robust and informative.

The PRSKB simplifies access to data required for polygenic risk score calculations. No other tool includes a centralized online database and command line interface that allow users to simultaneously query thousands of studies. We anticipate that the PRSKB will enhance the role of polygenic risk scores in future genetic studies of complex disease and trait heritability by streamlining the process to calculate polygenic risk scores.

Declarations

Competing Interests

JBM and JSKK cofounded The BYU Genetic Risk Assessment and PolyScores Reports, which is a commercial venture that calculates polygenic risk scores from consumer DNA tests.

Acknowledgements

We appreciate the contributions of Brigham Young University in supporting this research. This research is supported by RF1AG054052 (PI: Kauwe). Acknowledgement is made to the donors of Alzheimer's Disease Research, a program of the BrightFocus Foundation for support of this research through grant #A2020118F (PI: Miller).

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

References

1. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5–22, doi:10.1016/j.ajhg.2017.06.005 (2017).
2. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-d1012, doi:10.1093/nar/gky1120 (2019).
3. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genetics* **51**, 404–413, doi:10.1038/s41588-018-0311-9 (2019).
4. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* **45**, 1452–1458, doi:10.1038/ng.2802 (2013).
5. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature Genetics* **50**, 912–919, doi:10.1038/s41588-018-0152-6 (2018).
6. Hyde, C. L. *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nature Genetics* **48**, 1031–1036, doi:10.1038/ng.3623 (2016).
7. Zhao, W. *et al.* Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nature Genetics* **49**, 1450–1457, doi:10.1038/ng.3943 (2017).
8. Harold, D. *et al.* Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature Genetics* **41**, 1088–1093, doi:10.1038/ng.440 (2009).
9. Nikpay, M. *et al.* A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47**, 1121–1130, doi:10.1038/ng.3396 (2015).
10. Li, Z. *et al.* Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nature Genetics* **49**, 1576–1583, doi:10.1038/ng.3973 (2017).

11. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nature Reviews Cancer* **17**, 692–704, doi:10.1038/nrc.2017.82 (2017).
12. Liang, B., Ding, H., Huang, L., Luo, H. & Zhu, X. GWAS in cancer: progress and challenges. *Molecular genetics and genomics: MGG* **295**, 537–561, doi:10.1007/s00438-020-01647-z (2020).
13. *Cancer Epidemiology Biomarkers & Prevention* **27**, 363, doi:10.1158/1055-9965.EPI-16-0794 (2018).
14. Matoba, N. *et al.* GWAS of smoking behaviour in 165,436 Japanese people reveals seven new loci and shared genetic architecture. *Nature Human Behaviour* **3**, 471–477, doi:10.1038/s41562-019-0557-y (2019).
15. Walters, R. K. *et al.* Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nature Neuroscience* **21**, 1656–1669, doi:10.1038/s41593-018-0275-1 (2018).
16. Pal, L. R., Yu, C.-H., Mount, S. M. & Moult, J. Insights from GWAS: emerging landscape of mechanisms underlying complex trait disease. *BMC Genomics* **16 Suppl 8**, S4-S4, doi:10.1186/1471-2164-16-S8-S4 (2015).
17. Hirschhorn, J. N. Genomewide association studies–illuminating biologic pathways. *The New England journal of medicine* **360**, 1699–1701, doi:10.1056/NEJMp0808934 (2009).
18. Hampe, J. *et al.* A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat Genet* **39**, 207–211, doi:10.1038/ng1954 (2007).
19. Cao, C. & Moult, J. GWAS and drug targets. *BMC Genomics* **15 Suppl 4**, S5-S5, doi:10.1186/1471-2164-15-S4-S5 (2014).
20. Kauppi, K. *et al.* Revisiting Antipsychotic Drug Actions Through Gene Networks Associated With Schizophrenia. *Am J Psychiatry* **175**, 674–682, doi:10.1176/appi.ajp.2017.17040410 (2018).
21. Zhang, J. *et al.* Use of genome-wide association studies for cancer research and drug repositioning. *PLoS One* **10**, e0116477-e0116477, doi:10.1371/journal.pone.0116477 (2015).
22. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols* **15**, 2759–2772, doi:10.1038/s41596-020-0353-1 (2020).
23. Clyde, D. Making the case for more inclusive GWAS. *Nature Reviews Genetics* **20**, 500–501, doi:10.1038/s41576-019-0160-0 (2019).
24. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* **100**, 635–649, doi:10.1016/j.ajhg.2017.03.004 (2017).
25. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications* **10**, 3328, doi:10.1038/s41467-019-11112-0 (2019).
26. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications* **10**, doi:10.1038/s41467-019-11112-0 (2019).
27. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *New England Journal of Medicine* **375**, 655–665, doi:10.1056/nejmsa1507092 (2016).

28. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518, doi:10.1038/s41586-019-1310-4 (2019).

29. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med* **12**, 44, doi:10.1186/s13073-020-00742-5 (2020).

30. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genetics* **45**, 400–405, doi:10.1038/ng.2579 (2013).

31. Torkamani, A. & Topol, E. Polygenic Risk Scores Expand to Obesity. *Cell* **177**, 518–520, doi:https://doi.org/10.1016/j.cell.2019.03.051 (2019).

32. Jia, G. *et al.* Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers. *JNCI Cancer Spectrum* **4**, doi:10.1093/jncics/pkaa021 (2020).

33. Choi, J., Jia, G., Wen, W., Long, J. & Zheng, W. Evaluating polygenic risk scores in assessing risk of nine solid and hematologic cancers in European descendants. *International journal of cancer* **147**, 3416–3423, doi:https://doi.org/10.1002/ijc.33176 (2020).

34. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* **50**, 1219–1224, doi:10.1038/s41588-018-0183-z (2018).

35. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature reviews. Genetics* **19**, 581–590, doi:10.1038/s41576-018-0018-x (2018).

36. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet* **104**, 21–34, doi:10.1016/j.ajhg.2018.11.002 (2019).

37. Kuchenbaecker, K. B. *et al.* Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. *Journal of the National Cancer Institute* **109**, djw302, doi:10.1093/jnci/djw302 (2017).

38. Chen, J. *et al.* Polygenic Risk Scores for Subtyping of Schizophrenia. *Schizophrenia Research and Treatment* 2020, 1638403, doi:10.1155/2020/1638403 (2020).

39. Fritsche, L. G. *et al.* Exploring various polygenic risk scores for skin cancer in the phenomes of the Michigan genomics initiative and the UK Biobank with a visual catalog: PRSWeb. *PLoS genetics* **15**, e1008202 (2019).

40. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *The American Journal of Human Genetics* **104**, 21–34, doi:https://doi.org/10.1016/j.ajhg.2018.11.002 (2019).

41. Aminoff, S. R. *et al.* Polygenic risk scores in bipolar disorder subgroups. *Journal of Affective Disorders* **183**, 310–314, doi:https://doi.org/10.1016/j.jad.2015.05.021 (2015).

42. Power, R. A. *et al.* Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nature Neuroscience* **18**, 953–955, doi:10.1038/nn.4040 (2015).

43. Cross-Disorder Group of the Psychiatric Genomics, C. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet (London, England)* **381**, 1371–

1379, doi:10.1016/S0140-6736(12)62129-1 (2013).

44. International Schizophrenia, C. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752, doi:10.1038/nature08185 (2009).
45. Goris, A. *et al.* No evidence for shared genetic basis of common variants in multiple sclerosis and amyotrophic lateral sclerosis. *Human molecular genetics* **23**, 1916–1922, doi:10.1093/hmg/ddt574 (2014).
46. Agerbo, E. *et al.* Polygenic Risk Score, Parental Socioeconomic Status, Family History of Psychiatric Disorders, and the Risk for Schizophrenia: A Danish Population-Based Study and Meta-analysis. *JAMA psychiatry* **72**, 635–641, doi:10.1001/jamapsychiatry.2015.0346 (2015).
47. Mullins, N. *et al.* Polygenic interactions with environmental adversity in the aetiology of major depressive disorder. *Psychological medicine* **46**, 759–770, doi:10.1017/s0033291715002172 (2016).
48. Hindy, G. *et al.* Cardiometabolic Polygenic Risk Scores and Osteoarthritis Outcomes: A Mendelian Randomization Study Using Data From the Malmö Diet and Cancer Study and the UK Biobank. *Arthritis & Rheumatology* **71**, 925–934, doi:https://doi.org/10.1002/art.40812 (2019).
49. Richardson, T. G., Harrison, S., Hemani, G. & Davey Smith, G. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *eLife* **8**, e43657, doi:10.7554/eLife.43657 (2019).
50. Fontanillas, P. *et al.* Disease risk scores for skin cancers. *Nature Communications* **12**, 160, doi:10.1038/s41467-020-20246-5 (2021).
51. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* **8**, doi:10.1093/gigascience/giz082 (2019).
52. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575, doi:10.1086/519795 (2007).
53. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74, doi:10.1038/nature15393 (2015).
54. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209, doi:10.1038/s41586-018-0579-z (2018).
55. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47**, D1005-D1012, doi:10.1093/nar/gky1120 (2018).
56. Magno, R. & Maia, A. T. gwasrapidd: an R package to query, download and wrangle GWAS catalog data. *Bioinformatics (Oxford, England)* **36**, 649–650, doi:10.1093/bioinformatics/btz605 (2020).
57. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome research* **9**, 677–679 (1999).
58. Slatkin, M. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetics* **9**, 477–485, doi:10.1038/nrg2361 (2008).

59. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, doi:10.1186/s13742-015-0047-8 (2015).

60. Guo, Y. *et al.* Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* **109**, 83–90, doi:https://doi.org/10.1016/j.ygeno.2017.01.005 (2017).

61. Privé, F., Vilhjálmsson, B. J., Aschard, H. & Blum, M. G. B. Making the Most of Clumping and Thresholding for Polygenic Scores. *The American Journal of Human Genetics* **105**, 1213–1221, doi:https://doi.org/10.1016/j.ajhg.2019.11.001 (2019).

62. Wray, N. R. *et al.* Research review: Polygenic methods and their application to psychiatric traits. *Journal of child psychology and psychiatry, and allied disciplines* **55**, 1068–1087, doi:10.1111/jcpp.12295 (2014).

63. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research* **17**, 1520–1528, doi:10.1101/gr.6665407 (2007).

64. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752, doi:10.1038/nature08185 (2009).

65. Polderman, T. J. C. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics* **47**, 702–709, doi:10.1038/ng.3285 (2015).

66. Inc., P. T. (Plotly Technologies Inc., 2015).

67. Morris, J. C. The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology* **43**, 2412–2414, doi:10.1212/WNL.43.11.2412-a (1993).

68. Shapiro, S. S. & Wilk, M. B. An analysis of variance test for normality (complete samples)†. *Biometrika* **52**, 591–611, doi:10.1093/biomet/52.3-4.591 (1965).

69. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18**, 50–60, doi:10.1214/aoms/1177730491 (1947).

70. Trueger, N. S. *et al.* The altmetric score: a new measure for article-level dissemination and impact. *Annals of emergency medicine* **66**, 549–553 (2015).

71. Ellinghaus, D. *et al.* Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *The New England journal of medicine* **383**, 1522–1534, doi:10.1056/NEJMoa2020283 (2020).

72. Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. *The Lancet. Neurology* **15**, 174–184, doi:10.1016/s1474-4422(15)00338-5 (2016).

73. Shen, X. *et al.* A phenome-wide association and Mendelian Randomisation study of polygenic risk for depression in UK Biobank. *Nature Communications* **11**, 2301, doi:10.1038/s41467-020-16022-0 (2020).

74. Crouch, D. J. M. & Bodmer, W. F. Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proceedings of the National Academy of Sciences* **117**, 18924,

doi:10.1073/pnas.2005634117 (2020).

75. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753, doi:10.1038/nature08494 (2009).
76. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565–569, doi:10.1038/ng.608 (2010).
77. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427, doi:10.1038/nature13595 (2014).
78. Leonenko, G. *et al.* Genetic risk for alzheimer disease is distinct from genetic risk for amyloid deposition. *Annals of neurology* **86**, 427–435, doi:10.1002/ana.25530 (2019).
79. Vijayan, M. & Reddy, P. H. Stroke, Vascular Dementia, and Alzheimer’s Disease: Molecular Links. *Journal of Alzheimer's Disease* **54**, 427–443, doi:10.3233/JAD-160527 (2016).
80. Lucke-Wold, B. P. *et al.* Common mechanisms of Alzheimer's disease and ischemic stroke: the role of protein kinase C in the progression of age-related neurodegeneration. *Journal of Alzheimer's disease: JAD* **43**, 711–724, doi:10.3233/jad-141422 (2015).
81. Dong, S., Maniar, S., Manole, M. D. & Sun, D. Cerebral Hypoperfusion and Other Shared Brain Pathologies in Ischemic Stroke and Alzheimer’s Disease. *Translational Stroke Research* **9**, 238–250, doi:10.1007/s12975-017-0570-2 (2018).
82. Rahman, M. R. *et al.* Discovering Biomarkers and Pathways Shared by Alzheimer’s Disease and Ischemic Stroke to Identify Novel Therapeutic Targets. *Medicina* **55**, doi:10.3390/medicina55050191 (2019).
83. Mills, M. C. & Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nature Genetics* **52**, 242–243, doi:10.1038/s41588-020-0580-y (2020).
84. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20**, 467–484, doi:10.1038/s41576-019-0127-1 (2019).

Figures

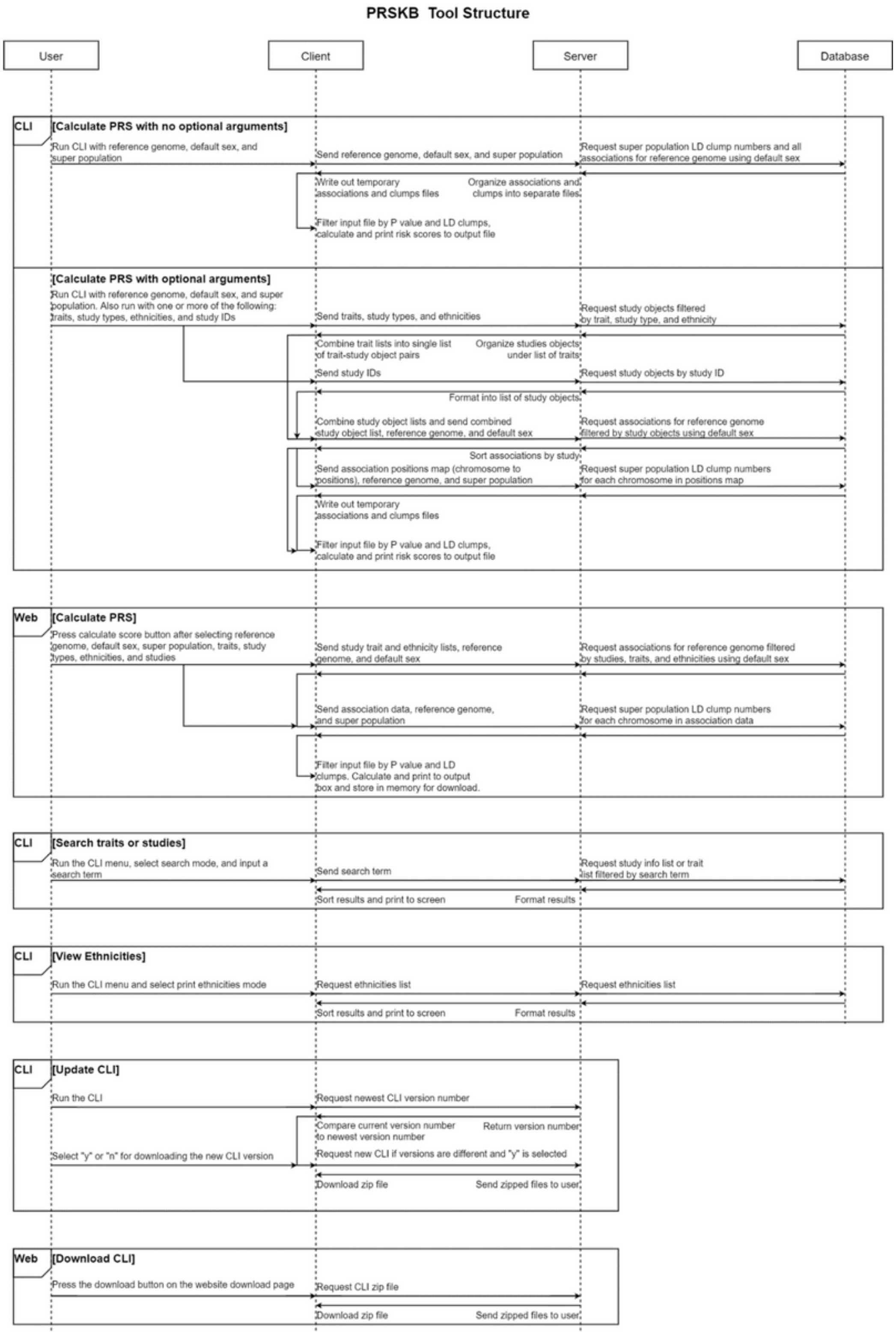
PRSKB Tool Structure
User
Client
Server
Database
CLI
[Calculate PRS with no optional arguments]
Run CLI with reference genome, default sex, and super population
Send reference genome, default sex, and super population
Request super population LD clump numbers and all associations for reference genome using default sex
Write out temporary associations and clumps files
Organize associations and clumps into separate files
Filter input file by P value and LD clumps, calculate and print risk scores to output file
[Calculate PRS with optional arguments]
Run CLI with reference genome, default sex, and super population. Also run with one or more of the following: traits, study types, ethnicities, and study IDs
Send traits, study types, and ethnicities
Request study objects filtered by trait, study type, and ethnicity
Combine trait lists into single list of trait-study object pairs
Organize studies objects under list of traits
Send study IDs
Request study objects by study ID
Format into list of study objects
Combine study object lists and send combined study object list, reference genome, and default sex
Request associations for reference genome filtered by study objects using default sex
Sort associations by study
Send association positions map (chromosome to positions), reference genome, and super population
Request super population LD clump numbers for each chromosome in positions map
Write out temporary associations and clumps files
Filter input file by P value and LD clumps, calculate and print risk scores to output file
Web
[Calculate PRS]
Press calculate score button after selecting reference genome, default sex, super population, traits, study types, ethnicities, and studies
Send study trait and ethnicity lists, reference genome, and default sex
Request associations for reference genome filtered by studies, traits, and ethnicities using default sex
Send association data, reference genome, and super population
Request super population LD clump numbers for each chromosome in association data
Filter input file by P value and LD clumps. Calculate and print to output box and store in memory for download.
CLI
[Search traits or studies]
Run the CLI menu, select search mode, and input a search term
Send search term
Request study info list or trait list filtered by search term
Sort results and print to screen
Format results
CLI
[View Ethnicities]
Run the CLI menu and select print ethnicities mode
Request ethnicities list
Request ethnicities list
Sort results and print to screen
Format results
CLI
[Update CLI]
Run the CLI
Request newest CLI version number
Compare current version number to newest version number
Return version number
Select "y" or "n" for downloading the new CLI version
Request new CLI if versions are different and "y" is selected
Download zip file
Send zipped files to user
Web
[Download CLI]
Press the download button on the website download page
Request CLI zip file
Download zip file
Send zipped files to user

Figure 1

PRSKB Tool Structure. The PRSKB tool is composed of a client, a server, and a database. The user interacts with the client, which is either the web tool (https://prs.byu.edu), or the command-line interface (CLI). The client connects to the server that then retrieves and returns data from the PRSKB database to the client. Arrows in this diagram represent the flow of data. Boxes represent specific actions a PRSKB user can take and the icon in the top left of each box indicates the client type for each box.

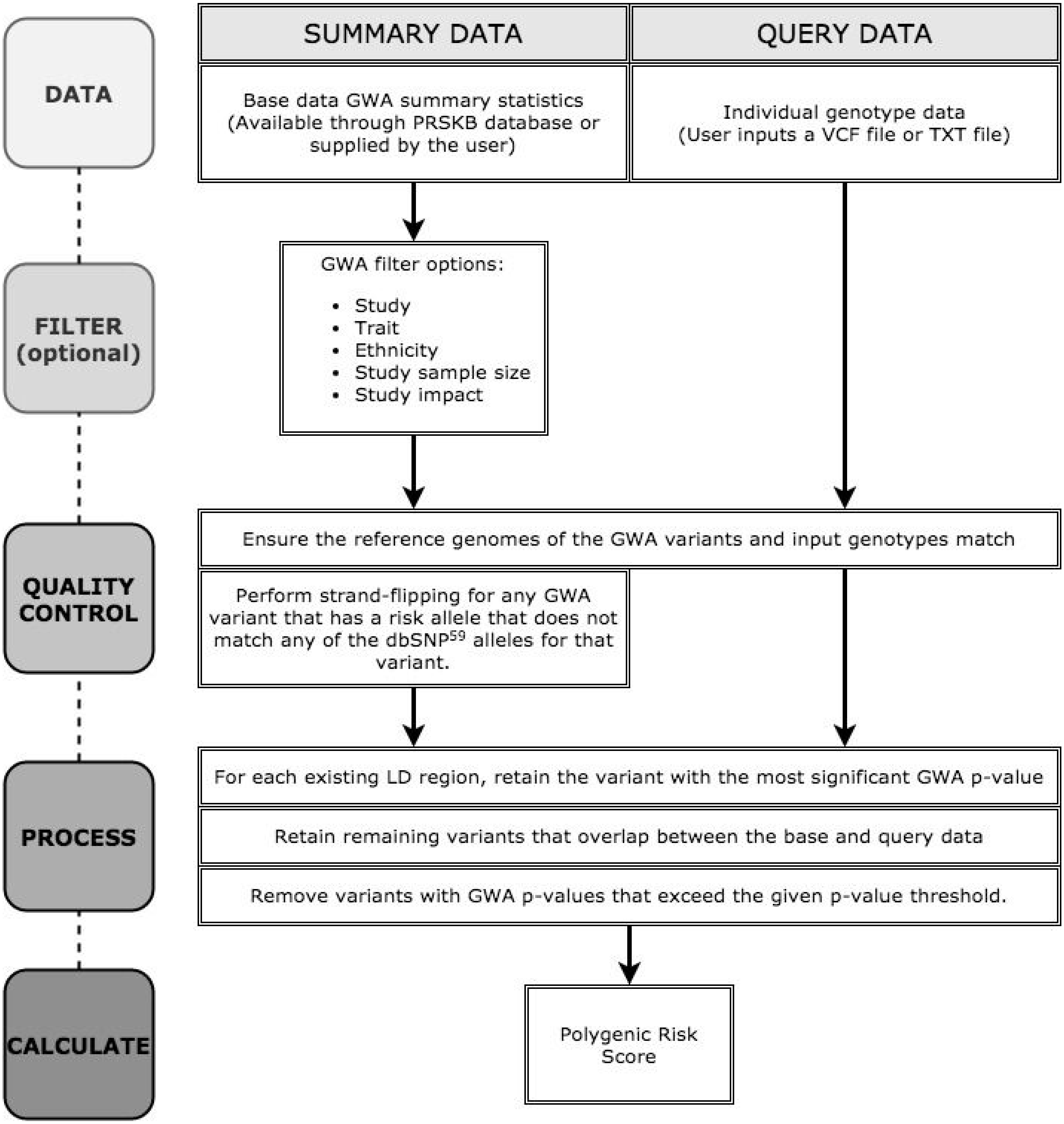
DATA
FILTER (optional)
QUALITY CONTROL
PROCESS
CALCULATE
SUMMARY DATA
QUERY DATA
Base data GWA summary statistics (Available through PRSKB database or supplied by the user)
Individual genotype data (User inputs a VCF file or TXT file)
GWA filter options:
• Study
• Trait
• Ethnicity
• Study sample size
• Study impact
Ensure the reference genomes of the GWA variants and input genotypes match
Perform strand-flipping for any GWA variant that has a risk allele that does not match any of the dbSNP59 alleles for that variant.
For each existing LD region, retain the variant with the most significant GWA p-value
Retain remaining variants that overlap between the base and query data
Remove variants with GWA p-values that exceed the given p-value threshold.
Polygenic Risk Score

Figure 2

Polygenic risk score workflow. The process follows the standards established by Choi, et al. 22

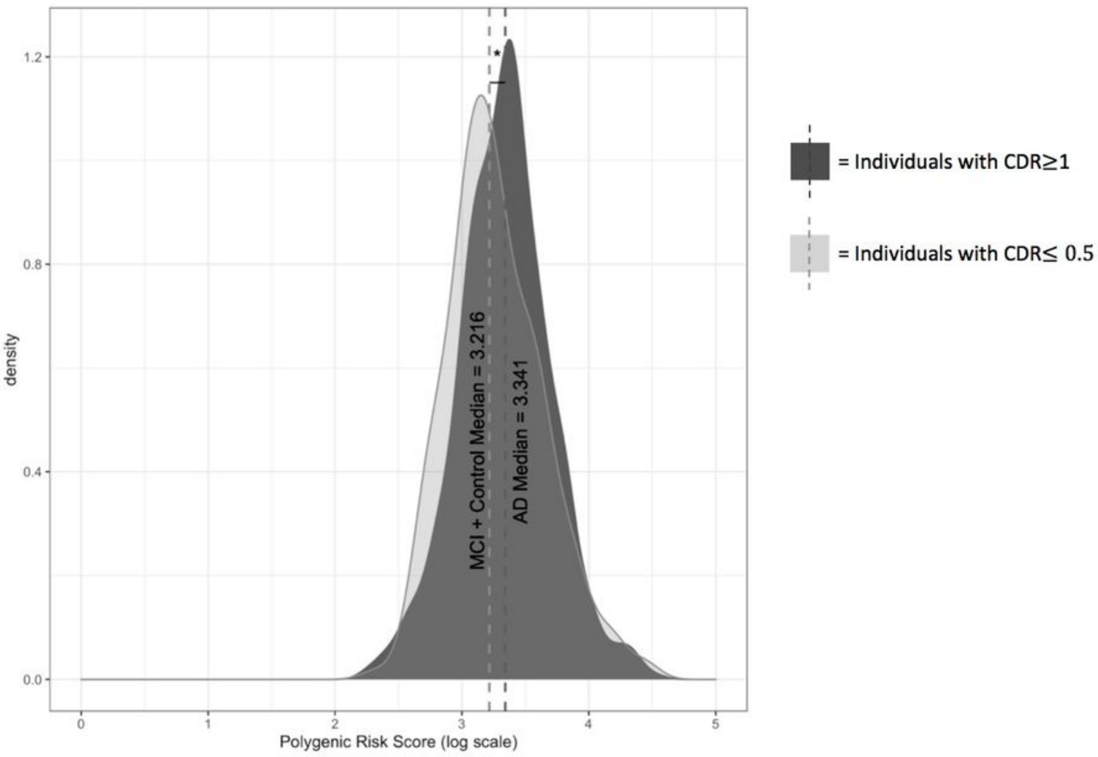

Figure 3

Alzheimer’s disease polygenic risk score distribution for ADNI participants with a CDR≥1 and a CDR≤0.5. A Mann-Whitney U test reveals that the median risk score in the individuals with a CDR≥1 is significantly higher than the median risk score in individuals with a CDR≤0.5 (U=67845, P=0.00332).

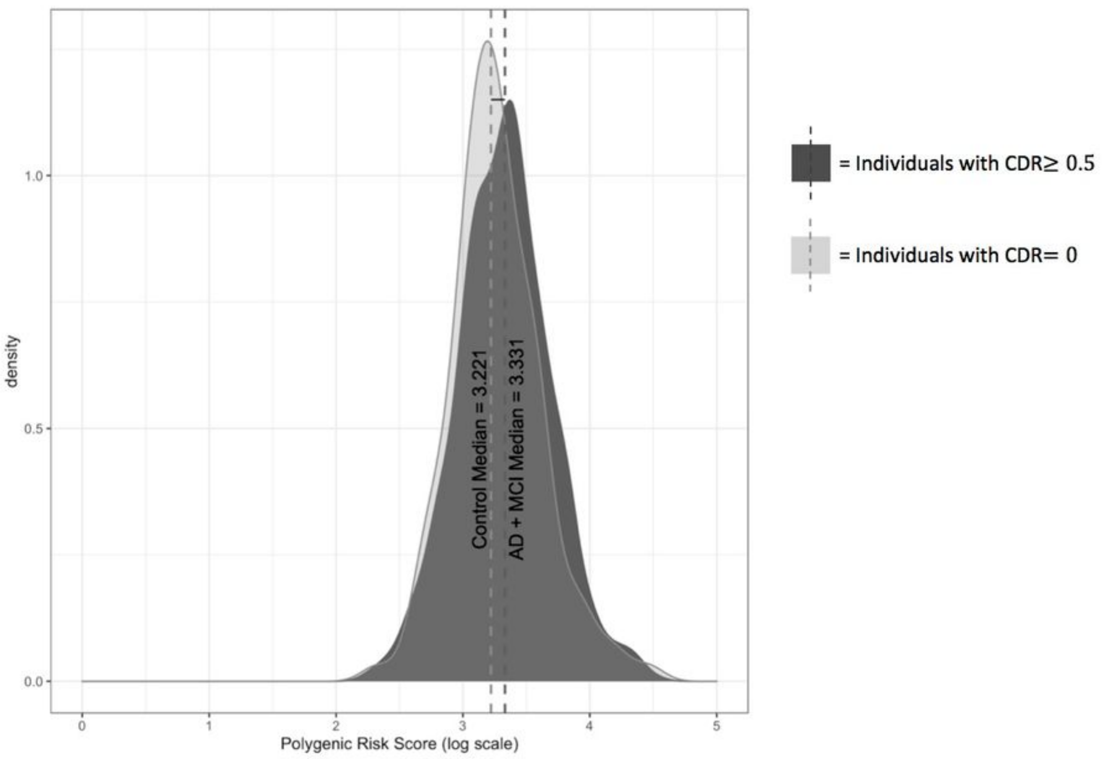

Figure 4

Alzheimer's disease polygenic risk score distribution for ADNI participants with a CDR≥0.5 and ADNI participants with a CDR=0. A Mann-Whitney U test reveals that the median risk score in the individuals with a CDR≥0.5 is not significantly different than the median risk score in individuals with a CDR=0 (U=42191; p=0.04273).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PRSKBsupplementaltables.pdf
- SupplementaryInformation.docx