DRAGON-Data: A platform and protocol for integrating genomic and phenotypic data across large psychiatric cohorts

Leon Hubbard*[1], Amy J. Lynham*[1], Sarah Knott*[1], Jack F. G. Underwood[1], Richard Anney[1], Jonathan I. Bisson[1], Marianne.B.M van den Bree[1], Nick Craddock[1], Michael O'Donovan[1], Ian Jones[1],George Kirov[1], Kate Langley[1], Joanna Martin[1], Frances Rice[1], Neil Roberts[1], Anita Thapar[1], Michael J. Owen[1], Jeremy Hall[1], Antonio F. Pardiñas**[1], James T.R. Walters**[1]

Affiliations

1. MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK

* These authors contributed equally to the manuscript

** *Corresponding Authors*

Email: waltersjt@cardiff.ac.uk or pardinasa@cardiff.ac.uk

MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Hadyn Ellis Building, Maindy Road, Cardiff University, Cardiff, UK, CF24 4HQ

Abstract

Introduction:

Current psychiatric diagnoses, although heritable, have not been clearly mapped onto distinct underlying pathogenic processes. The same symptoms often occur in multiple disorders, and a substantial proportion of both genetic and environmental risk factors are shared across disorders. However, the relationship between shared symptomatology and shared genetic liability is still poorly understood. Well-characterised, cross-disorder samples are needed to investigate this matter, but currently few exist, and severe mental disorders are poorly represented in existing biobanking efforts. Purposely curated and aggregated data from individual research groups can fulfil this unmet need, resulting in rich resources for psychiatric research.

Methods and analyses:

As part of the Cardiff MRC Mental Health Data Pathfinder, we have curated and harmonised phenotypic and genetic information from 15 studies within the MRC Centre for Neuropsychiatric Genetics and Genomics to create a new data repository, DRAGON-DATA. To date, DRAGON-DATA includes over 45,000 individuals: adults or children with psychiatric diagnoses, affected probands with family members and individuals who carry a known neurodevelopmental copy number variant (ND-CNV). We have processed the available phenotype information to derive core variables that can be reliably analysed across groups. In addition, all datasets with genotype information have undergone rigorous quality control, imputation, CNV calling and polygenic score generation.

Ethics and Dissemination:

DRAGON-DATA combines genetic and non-genetic information and is available as a resource for research across traditional psychiatric diagnostic categories. Its structure and governance follow standard UK ethical requirements (at the level of participating studies and the project as a whole) and conforms to principles reflected in the EU data protection scheme (GDPR). Algorithms and pipelines used for data harmonisation are currently publicly available for the scientific community, and an appropriate data sharing protocol will be developed as part of ongoing projects (DATAMIND) in partnership with HDR UK.

Introduction

The value of collaboration and data sharing is well recognised within the medical community and is one of the hallmarks of what has been called "the fourth age of research", in which the pace of discovery has accelerated and international platforms for studying multifactorial problems have been built[1 2]. The aggregation of data from individual research groups not only maximises the utility of individual datasets and minimises demands on participants, but enables the joint analyses of complex data that can lead to incremental advances in elucidating disease aetiology[3]. Within major psychiatric and neurodevelopmental conditions, few truly novel pharmacological treatments have been developed for several decades, with the noteworthy exceptions of ketamine for depression[4] and atomoxetine for ADHD[5]. Worryingly, many major pharmaceutical companies are decreasing their research efforts and investment in this area[6]. This apparent stagnation in progress is the result of a lack of understanding of the pathogenesis of these conditions, which hinders the identification of novel targets for drug discovery[7], and also the limitations of current diagnostic categories in defining mechanistically discrete disorders[8]. A route to address these limitations involves integrating biological data at scale and across, rather than within, diagnostic classifications[9 10]. Research conducted in this manner can explore the aetiological and biological commonalities between diagnoses revealed by genetic studies[11], accelerating discovery on complex disorders and informing novel therapeutic strategies, pharmacological and non-pharmacological, firmly grounded in biology[12].

Recent large-scale studies have built on the hypothesis that psychiatric phenotypes do not always reflect distinct underlying pathogenic processes and that some genetic risk factors are shared between neuropsychiatric disorders[13-15]. This echoes the widely acknowledged clinical observation that many symptoms are features of multiple disorders and that patients often challenge current diagnostic classifications by presenting with characteristics of more than one disorder[16]. What is currently not known, however, is to what extent this distribution of cross-disorder symptoms is related to the shared genetic liability between neurodevelopmental conditions[15 17]. Commonalities in genetic risk factors might help identify a shared underlying biology, but this line of inquiry cannot be pursued without well-characterised cross-disorder samples, scarce even within large international consortia. In fact, it has been explicitly suggested that the majority of samples used in published genetic discovery studies have not been collected with the required amount of phenotypic data necessary to advance diagnostics, stratification and treatment[18]. Thus, many research groups have directed their efforts to access resources with large amounts of routinely collected data, such as population biobanks and electronic health record systems, from which rich phenotypic data can be derived[18-20]. However, some common limitations of these include selection biases and underrepresentation of clinically severe disorders[20 21]. These can be exemplified by a recent genetic study on 106,160 patients across four US healthcare systems, where only 522 individuals with a ICD-9/10 diagnosis of schizophrenia were included[22]. Such is a classic quandary in psychiatric genomics[23], in which the setup of research studies leads to either a large case sample with minimal phenotyping or an extensively phenotyped one with fewer individuals.

The *Digital Repository for Amalgamating GenOmic and Neuropsychiatric Data* (DRAGON-Data) was therefore established at Cardiff University as a means of developing a platform

where cross-disorder analyses of large well-phenotyped samples are possible. This approach integrates multiple existing case datasets with genetic, clinical, environmental, and developmental data. The focus on mental health across disorder boundaries and at scale aims to improve understanding of the pathophysiology of adult and child-onset neurodevelopmental and psychiatric disorders, providing opportunities to combine diagnosis-led and symptom-led research. DRAGON-Data shares a focus with previous initiatives to collate psychiatric phenotype data, which have included the Genetics of Endophenotypes of Neurofunction to Understand Schizophrenia (GENUS) consortium[24], the International Consortium for Schizotypy Research (ICSR)[25], the International 22q11.2 Deletion Syndrome Brain Behaviour Consortium (22q11.2DS IBBC)[26], the Psychosis Endophenotypes International Consortium[27], the Genes to Mental Health (G2MH) network, and ongoing efforts to collate phenotype data within the Psychiatric Genomics Consortium (PGC)[28]. However, all these projects have typically focused on a single disorder or group of closely related disorders, while DRAGON-Data seeks to integrate data from a range of psychiatric disorders across the symptomatology and developmental continua.

The current paper describes the formation of DRAGON-Data through the curation and harmonisation of phenotypic and genetic information across existing cohorts. This process has been informed by a series of legal and ethical considerations on the evolving landscape of individual-level data sharing, which is required to ensure the sustainability of this repository as a resource for current and future researchers. Therefore, the governance framework of DRAGON-Data is also described, which enables the access and reuse of its data in ways that align with confidentiality regulations and the ethics of participating studies.

Methods and Analysis

Studies included

Fifteen studies from the MRC Centre for Neuropsychiatric Genetics and Genomics at Cardiff University (MRC CNGG; https://www.cardiff.ac.uk/mrc-centre-neuropsychiatric-genetics-genomics) were included in this project. A summary of the studies can be found in **Table 1**. Each study had its own approved research ethics, whilst ethical approval for the curation and development of DRAGON-Data was obtained from Cardiff University's School of Medicine Research Ethics Committee (Ref: 19/72). The studies included participants who were adults with psychiatric disorders, children (defined as up to age 16 or age 18) with neurodevelopmental disorders, children of parents with psychiatric disorders, and both children and adult carriers of rare neurodevelopmental risk copy number variants (ND-CNVs).

Phenotypic data harmonisation strategy

The process of curating the phenotypic data is outlined in Figure 1. Initially, investigators from all studies completed a proforma detailing the data and types of measures available, including the study clinical interviews, rating scales and self-report questionnaires. All but one of the studies included a structured clinical interview, and thus consistent symptom-level data were available (Error! Reference source not found.), along with a detailed phenotype data.

We compared all the variables to identify overlaps and resolve situations where the same information might have been differently labelled across studies. We also defined a core set of variables (**Table 2**), focused on information relevant and applicable to cross-disorder research. A primary consideration for including a variable among this core set was whether it was collected as part of the National Centre for Mental Health (NCMH) research programme. The NCMH is a Welsh Government-funded research centre that investigates neurodevelopmental, psychiatric and neurodegenerative disorders across the lifespan. Its cohort is the largest sample with phenotype data available to us, and a cross-disorder resource in itself[29]. As NCMH is still being expanded by recruitment of participants, maximising its compatibility with DRAGON-Data was desired. Additionally, every core variable was required to be available in at least half the current datasets, taking into consideration that some data might be specific to child or adult cohorts. Variables that were not available in NCMH and were present in less than half the studies were only included if they could be derived from existing data to achieve the representation threshold.

Challenges of harmonising phenotypic data

Measuring and rating psychopathology

The individual studies that form DRAGON-Data were designed using standard protocols for psychiatric research, and collected similar phenotypic data. However, they also used a range of different interviews, rating scales and questionnaires.This creates well-known challenges for data harmonisation[30 31]. In general, it should be noted that caution has to be exercised when amalgamating data from different studies even when these claim to use the same measures. Potential differences can include:

- Versioning: Measures can differ considerably between versions, with items being added or removed and definitions changing.
- Rating definitions: Ordinal scales can be named (e.g. 1="mild", 2="moderate", etc) resulting in a categorical or integer variable depending on study protocol. Some scales (e.g. OPCRIT[32]) can include items for which decimal point rating is acceptable, which could be transformed into continuous variables.
- Rating timeframes: Symptom and event data can be evaluated over different timeframes spanning weeks, months or years; and recorded as current, worst or lifetime occurrences. When integrating adult and childhood studies, it should be considered that events defined for the "lifetime" are not directly comparable due to intrinsic differences in this period of assessment. Measures that evaluate personality and behavioural traits might also not be completely consistent given the changes in these throughout the lifetime[33].
- Sources of information: A difference between adult and childhood studies is that the latter is more likely to use multiple informants (participants, their siblings, parents and teachers). Harmonising all these reports can be difficult and might also require a prior compatibility assessment[34].

The considerations above apply to individual studies, but they can add particular difficulty to reflect complex outcomes in a larger harmonised dataset. As an example, we highlight the different ratings of suicidal ideation across the DRAGON-Data studies (**Table 3**). Note that

these studies differed in whether they considered single versus multiple suicide attempts, duration of suicidal ideation or seriousness of attempts. This is likely to reflect the existence of different definitions of suicidal behaviour used in different research contexts[35 36], and illustrates one of the challenges that can be faced when merging data from different studies.

Sampling from the Population

Recruitment strategies and inclusion criteria can affect the characteristics of the samples, creating differences between them and making them unrepresentative of the population from which they are drawn. It has been suggested that participants enrolled in research studies of serious mental illness display better functional outcomes than are typical for those with the disorders in the wider population[37], when compared against naturalistic samples from outpatient services[38]. Population cohort studies also suggest that those with more severe psychopathology and higher genetic loading for psychiatric disorder are more likely to drop out, leading to underrepresentation particularly in longitudinal samples[39]. Media used to approach these participants also play a role in the sample characteristics, with internet-based recruitment engaging larger proportions of highly-educated female individuals but also those from ethnic minorities[40 41]. For most studies in DRAGON-Data, recruitment was based on clinically ascertained, prevalent cases and therefore are likely to have over-sampled participants with severe, chronic illness and under-sampled individuals who recovered and/or were discharged from services. Additionally, in common mental health conditions such as depression and anxiety, this might also over-represent women who are more likely to access help than affected males[42]. A special case in terms of sample composition also concerns the DEFINE, ECHO and IMAGINE studies, which specifically focused on carriers of ND-CNVs. Including these samples has important implications for research examining genotype-phenotype associations in the combined dataset, as improperly accounting for their genotype-led recruitment might bias calculations on the prevalence of genetic or environmental risk factors. However, they are important to integrate as they also enable comparative research into the role of these in people with and without highly penetrant genetic variants[43].

Study Protocol

Samples were recruited following longitudinal and cross-sectional designs. The existence of a follow-up period in longitudinal studies establishes a temporal order for symptom and event measures, which provides another level of detail over the broader definitions found in cross-sectional designs. The cross-sectional studies collected a mixture of current, worst episode and lifetime symptom measures. As it has been previously described in the context of causal inference[44], it is not advisable to combine longitudinal measures into or with “lifetime ever” variables, since this assumes that the events they reflect did not occur outside of the study assessment periods. Other issues that can affect the compatibility of different designs are attrition (in longitudinal studies), participant issues in completing assessments (e.g. length of time required) and the mode in which the study was conducted. Within DRAGON-Data most studies were conducted face to face with participants before the onset of the COVID-19 pandemic, but also utilised telephone interviews, postal questionnaires and online data collection. This could affect how questions are interpreted and in turn, the likelihood and

content of participants' responses. In addition, there is evidence that participants may be more willing to disclose sensitive information in some settings than others[45 46].

Diagnosis

Due to the different focus of individual DRAGON-Data studies, there were differences in the ways that diagnoses were made. Most studies used standardised interviews and medical records (where available) to derive consensus research diagnoses, with CLOZUK validating their ascertainment (based on intake of the antipsychotic clozapine) against research interviews[47]. The NCMH population sample used self-report, asking participants to report diagnoses that they had been given by a health professional. This is an approach taken by other large studies such as the UK Biobank[48]. While data obtained via self-reports can be of poorer resolution than that from a structured interview, this approach has the advantage of allowing faster recruitment of larger samples[49]. The accuracy of self-report diagnoses needs also to be considered, which may differ by diagnosis. Self-reported diagnoses of specific, chronic mental health conditions that require involvement with secondary psychiatric services, such as schizophrenia, may be more accurate than reports of common mental health conditions, such as depression, that typically encompass a wide range of presentations and can be diagnosed and treated in a variety of health settings. This can introduce variability in defining phenotypes with impacts on study results. Research attempting to estimate the heritability of depressive disorders using inconsistent diagnostic criteria classically demonstrated this[50]; and recent work employing samples with broad, self-report definitions of depression to identify genetic risk loci have also resulted in signals that are not specific to this condition[51]. To ameliorate these problems, the studies included in DRAGON-Data have focused on categorical diagnoses rated according to the Diagnostic and Statistical Manual of Mental Disorders (DSM) or International Statistical Classification of Diseases (ICD) criteria. While these are standard criteria, it has been proposed that a better approach to diagnostic classification may be to focus on dimensional measures of psychopathology, such as the National Institute for Mental Health's Research Domain Criteria (RDoC[52]). This approach may be adopted in the future as it could facilitate combining datasets to conduct cross-disorder research, given that many symptoms overlap diagnostic boundaries, such as the overlapping mood and psychotic symptoms observed in both schizophrenia and bipolar disorder[8].

Key Recommendations

Based on our experience developing DRAGON-Data, we suggest some recommendations for the harmonisation and analysis of clinical data:

- Consider the broad research questions that can be addressed with the creation of a clinical database. Consult with principal investigators and field researchers to identify the variables that will be needed to address these aims.
- Identify measures (e.g., questionnaires and interviews) that are in common across the datasets included. These measures may be easier to harmonise for analysis, though the factors outlined above should be considered to ensure comparability.
- Record accurate information about each study variable including measure used, version number, rating definitions, rating timeframe and source of information. This aids in the identification of comparable variables.

- Where new (secondary) variables have been derived from others, and are designed to be comparable, information should be recorded about the (primary) variables used from each study to derive those secondary variables.
- A comprehensive data dictionary should accompany the database that incorporates the information outlined above. At a minimum, each variable should have recorded: name, description, definition and coding of missing values. Within the data dictionary, variables should be highlighted if they are in common across the datasets, as these may be suitable to analyse together. It is noteworthy that this curation and creation of dictionaries may often need to occur after the data collection, so researchers and funders should allow sufficient staff resources for the accurate completion of this task.
- Include basic demographic information to evaluate the representativeness of the sample, including age range, sex, ethnicity and education.
- Datasets do not need to be combined into a single data file. A database that houses the datasets and allows an easy combination of selected studies and variables avoids the need for a single, large-scale dataset and minimises the computational requirements for the querying and extraction of data.

Genetic data harmonisation strategy

Format and genome assembly standardisation

We developed an in-house genotype quality control (QC) pipeline to facilitate standardised procedures for all aspects of genetic analysis (**Figure 2**), available at https://github.com/CardiffMRCPathfinder/GenotypeQCtoHRC. The pipeline begins with conversion of genotype data into binary PLINK format[53 54]. Genotyping platform was inferred by comparing chromosome and basepair positions of the genotypes on each dataset and 166 array manifests[55]. Across the datasets in DRAGON-Data, Illumina chips are by far the most common (**Table 1**).

To maximise the number of SNPs available for imputation, we performed alignment of local genotype data against the Haplotype Reference Consortium (HRC) panel v.1.1[56] using Genotype Harmoniser v1.42[57]. Genotype Harmoniser is a Java-based application that compares SNP information in the user data against a reference dataset such as an imputation panel. Where discordant SNP information is present, for example due to allele mismatches, strand flips or different SNP identifiers, the user genotype data is updated to match that of the reference panel. We have observed that differences in genome build between the original and reference dataset result in Genotype Harmoniser discarding large numbers (e.g. more than 50%) of the original SNPs. If present, instances of this behaviour are flagged by our pipeline and solved via a local implementation of the widely-used Liftover Tool[58] to retrieve physical coordinates in the appropriate b37/hg19 format.

Sex-based quality control

We performed checks for discordant phenotypic and biological sex using the "sex-check" function in PLINK v1.9. This function is reliant on the presence of at least one sex chromosome. Discordant findings in the absence of complementary information from the

individual (e.g. a disclosure of gender transitioning) are suggestive of either a sample mix-up during genotyping or an inaccurately recorded phenotype. If no resolution can be reached these samples are excluded from further analysis. Where no sex information is present in the original dataset, the sample is retained. If genotype calls from both sex chromosomes are present, call rates at the Y chromosome are used to assess the presence of individuals with sex-linked chromosomal disorders such as Turner (X0) or Klinefelter (XXY) syndromes[59]. Individuals with suggestive sex-linked chromosomal disorders are flagged for further investigation.

Call-rate quality control

We removed SNPs with low call rates (<0.95), individuals with low genotyping rates (<0.95), markers that fail the Hardy-Weinberg Equilibrium test (mid-p<10^{-6}) and those with a minor allele frequency (MAF) < 0.01. Duplicated individuals were removed unless they belong to known monozygotic twin pairs; however, first degree relatives are retained for studies with trio or family designs. This is the final step of the pre-imputation QC. Afterwards genotypes are converted to VCF format using PLINK, sorted using vcftools v0.116[60] and compressed to .gz format.

Assessment of population structure

While not strictly part of a QC process, the generation of principal components (PCs) using genotype data is needed to identify and account for population and ancestral substructures that can bias the results of association studies[61]. Our pipeline addresses this by generating PCs using the GENESIS suite, implemented in R. Within it, the PC-AiR[62] function allows us to process both unrelated and family-based datasets, as it accounts for known or cryptic relatedness via the calculation of genotype relatedness matrices (GRMs). PCs generated by this method can readily be used to correct for population structure in regression-based analyses.

A more detailed ancestry analysis is also performed on each dataset, following a similar procedure to that described in Legge et al. 2019[63]. First the available SNPs are restricted to those on the set of 167 ancestry informative markers (AIMs) contained in the EUROFORGEN[64] and 55-AISNP[65] forensic panels, many of which are common across the different Illumina genotyping platforms. Afterwards, the dataset is merged with a public reference panel with known ancestries, a combination of the Human Genome Diversity Project (HGDP)[66] and South Asian Genome Project (SAGP)[67] datasets. This reference contains 1108 samples from 62 worldwide populations, which have been subdivided in 7 biogeographical ancestries[68] ("Subsaharan African", "North African", "European", "Southwest Asian", "East Asian", "Native American" and "Oceanian"). In order to perform the ancestry inference, a number of PCs, determined using the Tracy-Widom test for eigenvalues[61], are then derived solely on the reference panel, and a prediction model is trained using Fisher's Linear Discriminant Analysis algorithm. The samples with unknown ancestries are then "projected" onto the reference panel PCs[69], and their ancestry is estimated using the prediction model. At least 80% probability of a given ancestry is required to automatically assign an individual to it, though the admixture patterns of individuals not achieving this probability can still be manually examined.

Genotype imputation

The Michigan Imputation Server (MIS) is a cloud-based resource that facilitates haplotype pre-phasing and genotype imputation[70]. The MIS also houses the HRC panel, containing genotypes of over 60,000 individuals across multiple ancestral backgrounds[56]. There are substantial improvements in imputation quality using the HRC reference over 1000 genomes, particularly at lower MAF thresholds[71]. The MIS also performs some SNP quality control before phasing, including removal of SNPs if they contain irregular allele codes, duplicate IDs, indels, monomorphic SNPs, discordant alleles between the user and population reference panel alleles and low call rates of < 0.9. Though other options are available, our dataset is processed via Eagle v2.3 pre-phasing[72] and MiniMac3 imputation[70] using HRC v1.1 as the reference panel.

After genotype imputation, imputed data is stored in .vcf.gz format, with accompanying info files containing information about the quality of imputed variants. Data is converted into .pgen format using PLINK v2 and subsequently into standard .bed/.bim/.fam format. Specifically, we remove SNPs where individual genotype probabilities are < 0.9, MAF <1%, genotyping rate < 0.95 and hwe < 1E-4. SNPs can be extracted at various imputation quality thresholds (R2). A conversion to best-guess genotypes is also performed in PLINK v2, after applying imputation quality thresholds (INFO < 0.3).

Copy Number Variant Calling

Most of the samples in DRAGON-Data include raw genotype information, enabling us to perform copy number variant (CNV) calling. We developed an in-house CNV QC pipeline to facilitate standardised procedures for all aspects of this procedure (**Figure 3**), available at https://github.com/CardiffMRCPathfinder/NeurodevelopmentalCNVCalling.

First, we extract b-allele frequencies and logR ratios for each sample using Illumina Genome-Studio v2.05. CNV calling is performed using PennCNV v1.05 with genomic control correction[73]. CNVs are subsequently merged if the total distance between CNVs is less than 50% of their combined length. Appropriate PFB and GC content files are generated as recommended by PennCNV. Filters are applied to remove CNVs with QC fewer than 20 probes, less than 20KB in length or with confidence scores < 5. Individuals are excluded if they have more than 30 CNVs, large logR ratios > 0.35 or high or low wavefactor (less than -0.03 or greater than 0.03), however these parameters should be modified depending on the genotyping platform used.

Initially, CNVs called using this pipeline are cross-referenced against a list of 54 pathogenic CNVs known to confer increased risk of schizophrenia, autism, intellectual disability and major depressive disorder[74]. There are several advantages to prioritising these CNVs: First, they are typically large (>100KB) and are more reliably called across different genotyping platforms. Second, these CNVs are pleiotropic and lack complete penetrance for specific disorders meaning they are good candidates for investigating associations with psychiatric cross-disorder phenotypes.

Challenges of harmonising genetic data

Genotyping arrays and genomic assemblies

In DRAGON-Data, a variety of genotyping arrays were used both within and between studies. This presents challenges for merging and imputing datasets. All the genotyping arrays analysed have a large set of common variants (a "GWAS backbone"), with most differences due to the inclusion of custom markers tagging rare exonic variation. The accuracy of genotype imputation is improved with larger sample sizes, plateauing around 2,000 samples[75], though there must also be sufficient numbers of genotyped markers (at least 200,000 SNPs[76]) that overlap with the imputation reference panel after genotype quality control. We, therefore, grouped datasets that were genotyped on the same, or similar arrays. This resulted in four separate imputation batches for samples genotyped on the OmniExpress, PsychChip/Illumina HumanCoreExome, Illumina 610 Quad/Illumina HumanHap550 and Affymetrix5 platforms.

We observed substantial batch effects in the pairwise comparison of samples after undergoing routine QC. Further inspection of the data revealed this was caused by palindromic SNPs (AT/TA or CG/GC genotypes), which resulted in erroneous allele frequencies which differed across datasets when the minor allele frequency was high (> 0.4). This issue was only apparent after merging datasets, which mirrors the experience of the eMERGE consortium[77]. Removal of these SNPs resulted in the loss of obvious batch effects across the first 10 PCs tested.

Identifying duplicate samples

It is not uncommon for the same individual to be recruited into more than one psychiatric research study. Unless the individual voluntarily reports they have participated in a known existing study, this information would not be known to researchers in other groups. We identified 1909/41957 duplicate individuals (4.5%) across the entire dataset using genetic relatedness checks and retained the sample with the highest number of high quality imputed markers. In total,

Processing of public GWAS summary statistics

When performing genetic analyses such as polygenic risk scoring, LD score regression or other analyses, multiple GWAS summary statistics are required. Despite some proposals for standardisation[78-80], the output from GWAS software is still highly variable and lacks even consistent headings across individual studies. Processing of these files is thus not user-friendly, typically requiring manual curation, for example filtering by imputation quality, allele frequency or changing header names to match the required format of specific programs. To address these issues, we developed an R pipeline (summaRygwasqc) that automatically processes GWAS summary statistics files and performs quality control filtering, aligns SNP information against the HRC reference panel and converts summary data to a standardised format that is compatible with PRSICE2[81] , PRScs[82] and LDSC[83] (**Figure 4**). This code is available at https://github.com/CardiffMRCPathfinder/summaRygwasqc.

Key Recommendations

Based on our experience developing DRAGON-Data, we offer some recommendations for the amalgamation and analysis of genomic data across multiple studies:

1. Imputation should only be performed on samples that have been genotyped on the same array type, or where there is substantial SNP overlap after QC. Furthermore, when performing QC after imputation, removal of palindromic SNPs with high MAF (>0.4) is essential to minimising batch effects for samples genotyped on different arrays.
2. When analysing CNV data across arrays, due to potential differences in probe density and coverage, it is vital that plots such as those for b-allele frequency drift, number of CNVs called per individual and LogR ratio standard deviation are visually inspected to ensure the quality of the resulting calls.
3. Publicly available genome-wide association summary statistics should be examined, manually or through scripting, to ensure that their information can be processed in a coherent and standardised way in downstream analyses. At a minimum, most genomic analysis software requires a form of SNP name or identifier, chromosome number (CHR), basepair position (BP), allele code (A1/A2), association p-value and a metric for the association effect size (OR/logOR/beta/Z, which should always correspond with A1). Additional columns such as the allele frequency of A1, INFO/R2 imputation quality metrics and sample size columns can also be helpful.

The DRAGON-Data harmonised dataset

Table 2 displays an overview of the variables held by each study included in the final DRAGON-Data data freeze. A full list of the variables included in DRAGON-Data can be found in **Supplementary Table 1** although the exact variables included varied between studies. All the studies except CLOZUK included a semi-structured clinical diagnostic interview, most commonly the Schedule for Clinical Assessment in Neuropsychiatry (SCAN[84]) for adults and the Child and Adolescent Psychiatric Assessment (CAPA[85]) for children and adolescents. Twelve of the fifteen studies collected data on individual symptoms. The NCMH study includes a brief assessment that does not include questions about individual symptoms, although a subgroup of this sample (n=485) has completed more detailed interviews that include symptoms. The most common types of symptoms covered across all studies were depressive, manic and psychotic symptoms. Aside from symptoms, other variables with good coverage across studies were lifetime history of treatment (13/15), substance use (13/15) and history of suicidal ideation and attempts (12/15). The demographic characteristics of the studies are shown in **Supplementary Table 1**. The harmonised phenotype data is stored in a pseudonymised format within a secure database. There is an accompanying data dictionary cataloguing all available variables with names, descriptions and ratings and cross-referencing of comparable measures across the studies.

Joint genetic-phenotypic data analysis

All the DRAGON-Data data have been securely stored in HAWK, a high-performance computing (HPC) cluster supported by the Supercomputing Wales infrastructure[86], which comprises a network of 13,000 computer nodes distributed across four universities (Cardiff,

Swansea, Bangor and Aberystwyth). This system allows the backed-up storage of genetic and phenotypic files, and their secure access by authorised users. Analysts in charge of curating genetic or phenotypic data are by default part of a "core project team" with unrestricted access to the entire DRAGON-Data, while data-contributing researchers are granted access to their own raw and curated data for any purpose. Undertaking cross-disorder analyses is facilitated through a framework by which any curator or data-contributing researcher can send a structured analytic proposal to the board of investigators, who then decides whether to grant access to the relevant data on purely scientific grounds. This is modelled after successful international consortia such as the PGC[28], which in recent years has implemented responsible data sharing practices among hundreds of investigators.

There are two main approaches to analysing the data within DRAGON-Data: combining individual-level information from across the studies ("mega-analysis") or through meta-analysis. While the latter is relatively straightforward, jointly analysing all samples allows for a better assessment of heterogeneity in the data and can increase statistical power[87 88]. However, combining samples is particularly problematic for the phenotypic data, as it requires recoding or modifying the variables to be comparable across studies, which could include deriving latent variables through factor analysis. Data combined in this way can be difficult to interpret due to the differences between studies outlined in the previous sections, and it is important to address this variability in both analytic techniques and interpretation of the results. Important considerations are whether the individual study variables are measuring the same construct and whether any variables derived from these are measuring the same construct as the original data. Note that none of these limitations applies to the genetic data, as (carefully) combining samples with large numbers of overlapping SNPs is a common procedure that is known to maximise both the number of successfully imputed variants and their quality[75 89 90]. Thus, the suitability of a mega-analysis or meta-analysis approach for studies using DRAGON-Data should be decided based on the availability, characteristics and biases of the phenotypic data.

Outside of the data quality control pipelines, genetic analyses in DRAGON-Data can be undertaken using other consolidated tools, such as PLINK[53] or GCTA[91]. Responding to the rapid development of statistical methods to analyse complex phenotypes and "big data", an effort has been made to integrate DRAGON-Data with the highly customisable R framework, via the use of data importers such as *GWASTools*[92] and *bigsnpr*[93]. This allows using the approximately 1,700 tools currently offered by the Bioconductor suite[94] in a large-scale genome-wide setting, and facilitates applying complex analytic techniques such as mixed-model regression[95] and survival analysis[96]. Large-scale genomic storage solutions have not currently been implemented in DRAGON-Data, as the weak compression implemented in PLINK files and related formats allows for efficient querying of genotype data even in its imputed form[53 97]. However, these are active topics of research, and the upcoming development of the MPEG-G ISO standard will likely allow future data harmonisation initiatives to seamlessly incorporate whole-genome sequences[98].

Ethics and dissemination

Governance

For studies to be incorporated into DRAGON-Data, the lead principal investigator needed to confirm approval from their institutional ethics committee. The protection and confidentiality of participant data were of the utmost importance throughout the design of DRAGON-Data and a number of safeguards were put in place to ensure the security, integrity, accuracy and privacy of participant data. Firstly, in line with the required safeguards for processing special category data stipulated in the EU General Data Protection Regulation (GDPR; Article 89)[99], the principle of data minimisation was respected, with only limited individual-level data being requested from research groups. Furthermore, as a means of ensuring the confidentiality and privacy of participants, all data were pseudonymised, and no personal or phenotypic information that allowed individuals to be re-identified was retained. As genome-wide genetic information cannot effectively be anonymised without compromising its integrity[100], all researchers accessing it must explicitly state that they will not attempt participant re-identification.

This project was conducted in line with Cardiff University's Research Integrity and Governance Code of Practice, and ethical approval for the curation and development of the DRAGON-Data was obtained from Cardiff University's School of Medicine Research Ethics Committee (Ref: 19/72). As described above, procedural safeguards were put in place to ensure secure managed access to the dataset through the HAWK system, with the most privileges restricted to the "core analyst team". In addition, a process of oversight has been implemented for the approval of secondary research proposals, which are reviewed by the lead principal investigator of each contributing sample and must be approved before access to relevant, requested data can be granted. All genetic analyses carried out by secondary investigators also have to be carried out within the HAWK environment, which allows their monitoring and auditing to rapidly detect data misuses.

Challenges of data sharing partnerships

The organisational challenges faced by DRAGON-Data highlight that potential data sharing requirements should be considered, as much as reasonably possible, at the outset of any research study. Studies will benefit from having a data sharing policy in place prior to the collection of any data as a means of maximising the value of collected data, increasing transparency and ensuring responsible future sharing of data. This will depend on sharing with whom, and for what purpose. Consent processes have changed dramatically over the last 30 years and historical studies will not all have explicit consent on the data sharing practices that are more commonly included today[101]. In certain situations, additional ethical permission may be required for data sharing when the sample is historical and or individuals can no longer be contactable. Thus, data sharing without that explicit permission can only occur within certain circumscribed situations.

When obtaining consent for future research, researchers should aim to be as inclusive as possible and allow participants to provide their written informed consent for general areas of

research activity. In the context of broad consent, we would also advise the implementation of an oversight mechanism for the approval of future research studies. Participants entrust researchers to make reasonable decisions regarding future research on their behalf and the process of oversight adds further protection to participants, since not all future research uses can be predicted.

Dissemination

At present, DRAGON-Data has been designed as a way of maximising the present and future utility of data collected at the MRC CNGG during the last thirty years. Given the complexity of the data, particularly the phenotypic portion, the first cross-disorder analyses of DRAGON-Data have been carried out by members of the core analytic team and the participating investigator groups. Results of these analyses will be shared through Cardiff University online data repositories and communicated through standard scientific channels such as peer-reviewed publications. Ultimately, through adapting the PGC open science model[102] and taking advantage of the data-sharing frameworks supported by HDR UK, such as the DATAMIND Hub[103], the DRAGON-Data resource will be available for external investigators where individual study consent and ethics permit such data sharing. This will ensure compliance with the permissions and ethics of individual studies, and will be based on the secondary analysis principles detailed in the Governance section.

Table 1

Studies included in DRAGON-Data

Study	Reference	Main Diagnosis	Principal Investigator(s)	Genotyping Platform	N Genotyped (Post-QC)	Psychiatric Instruments Used	N Phenotyped (harmonised)
BDRN	[104]	Bipolar disorder	N. Craddock, I. Jones, L. Jones	Affymetrix5 OmniExpress PsychChip	4806 8035 1102	SCAN	6000
Bulgarian Trios							
Case-control data	[105]	Psychosis and mood disorders	G. Kirov	OmniExpress	806	SCAN	305
Family data*	[106]	Probands with psychosis and mood disorders and their families	G. Kirov	Affymetrix6	2119	SCAN	3084
CLOZUK	[47 107]	Treatment-resistant schizophrenia	J. T. R. Walters, M. Owen, M O'Donovan	OmniExpress	13743	None	16405
Cardiff COGS	[108]	Schizophrenia, psychosis or bipolar disorder	J. T. R. Walters, M. Owen	OmniExpress	997	SCAN	1301
CNV studies							
DEFINE	[109]	Confirmed ND-CNV carrier	J.Hall, D.Linden, M.B.M. van den Bree, M. Owen	PsychChip	971 (Number inclusive of	SCID PAS-ADD	125

ECHO IMAGINE	110 111	Confirmed ND-CNV carrier	M.B.M. van den Bree, J.Hall, D.Linden,M. Owen	PsychChip	ECHO and IMAGINE)	CAPA	963
EPAD*	112	Major depressive disorder (at least one affected parent)	F. Rice, A. Thapar	PsychChip	615	CAPA and SCAN	674
F-Series*	113	Psychosis and mood disorders	M. Owen	OmniExpress	749	SCAN	1022
DeCC/DeNt	114	Major depressive disorder	N. Craddock, L. Jones, C.Lewis, M.Owen	610 Quad	1346	SCAN	1504
NCMH	29	Any developmental or mental disorder	I. Jones (and others)	PsychChip	3352	SCAN (N=465) CAPS-5 PAS-ADD	16311
PTSD Registry	115	PTSD	J. Bisson, N. Roberts	PsychChip	325	SCID CAPS	325
SAGE*	116	ADHD	A. Thapar, M. O'Donovan, M.J. Owen, K. Langley, J. Martin	HumanHap550 PsychChip	2073*	CAPA	1132
Sib-Pairs	117	Schizophrenia	M. Owen	OmniExpress	918	SCAN	918

CAPA: Child and Adolescent Psychiatric Assessment; SCAN: Schedules for Clinical Assessment in Neuropsychiatry; SCID: Structured Clinical Interview for DSM-IV, *Includes family data and/or (trios).

Table 2

List of phenotypic variables included in DRAGON-Data

Variables Included	Number of studies	Number of participants
Symptoms		
Depression	12	15410
Mania	11	13906
Psychosis	9	12072
ADHD	4	2460
Anxiety	4	2478
Conduct disorders	4	2460
Autism	4	2460
PTSD	1	325
Treatment history	13	31164
Clinical / illness history		
Age of onset	10	29023
Hospital admissions	7	26372
Suicidal ideation	12	15410
Adverse life events	6	9594
Education	9	24790
Substance use	13	29997
Family history of psychiatric illness	8	21473
Physical health	11	27725
Functioning		
Standardised measure of functioning (e.g. Global Assessment Scale)	5	6260
Marital / relationship status	7	23290
Current occupation	7	25597
Cognitive function	7	5048

Number of participants refers to the number of data points available for each set of variables listed.

Table 3

Rating scales for suicidal ideation across the studies

Study	Suicidal Ideation: Rating Scale
CoMPaSS	0. Absent 1. Tedium Vitae 2. Suicidal Ideation 3. Attempt unlikely to result in death 4. Attempt likely to result in death 5. Multiple attempts likely to result in death
NCMH	0. Absent 1. Tedium vitae 2. Suicidal ideation 3. Attempt unlikely to result in death 4. Attempt likely to result in death 5. Multiple attempts unlikely to result in death 6. Multiple attempts likely to result in death
ECHO, IMAGINE, SAGE & EPAD (children only)	Binary variables (yes/no) covering: • Thoughts about death or suicide • Suicide attempts • Non-suicidal self-harm
EPAD (parents only)	Suicide attempt or self-harm: 1. Mild 2. Moderate 3. Severe
PTSD Registry	Question covers suicide attempts and self-harm in the context of borderline personality disorder: 1. Inadequate information 2. False or absent 3. Sub-threshold 4. Threshold or true
Sib-Pairs & F-series	0. None 1. 1 week duration or one attempt 2. 2 weeks duration 3. At least one month
Bulgarian Trios (family and case data)	0. Not present 1. Thoughts but no attempts 2. Attempt at suicide 3. Serious attempt

	4. Multiple serious attempt
BDRN	Suicidal ideation: 1. Yes 2. No 3. Unknown
DeCC and DeNt	1. Deliberately considered but no attempt 2. Injured self or made attempt but no serious harm 3. Suicide attempt resulting in serious harm 4. Suicide attempt designed to result in death 5. Uncertain

Note: No variable for suicidal ideation or attempts in DEFINE

Figure 1

Curation of phenotypic data

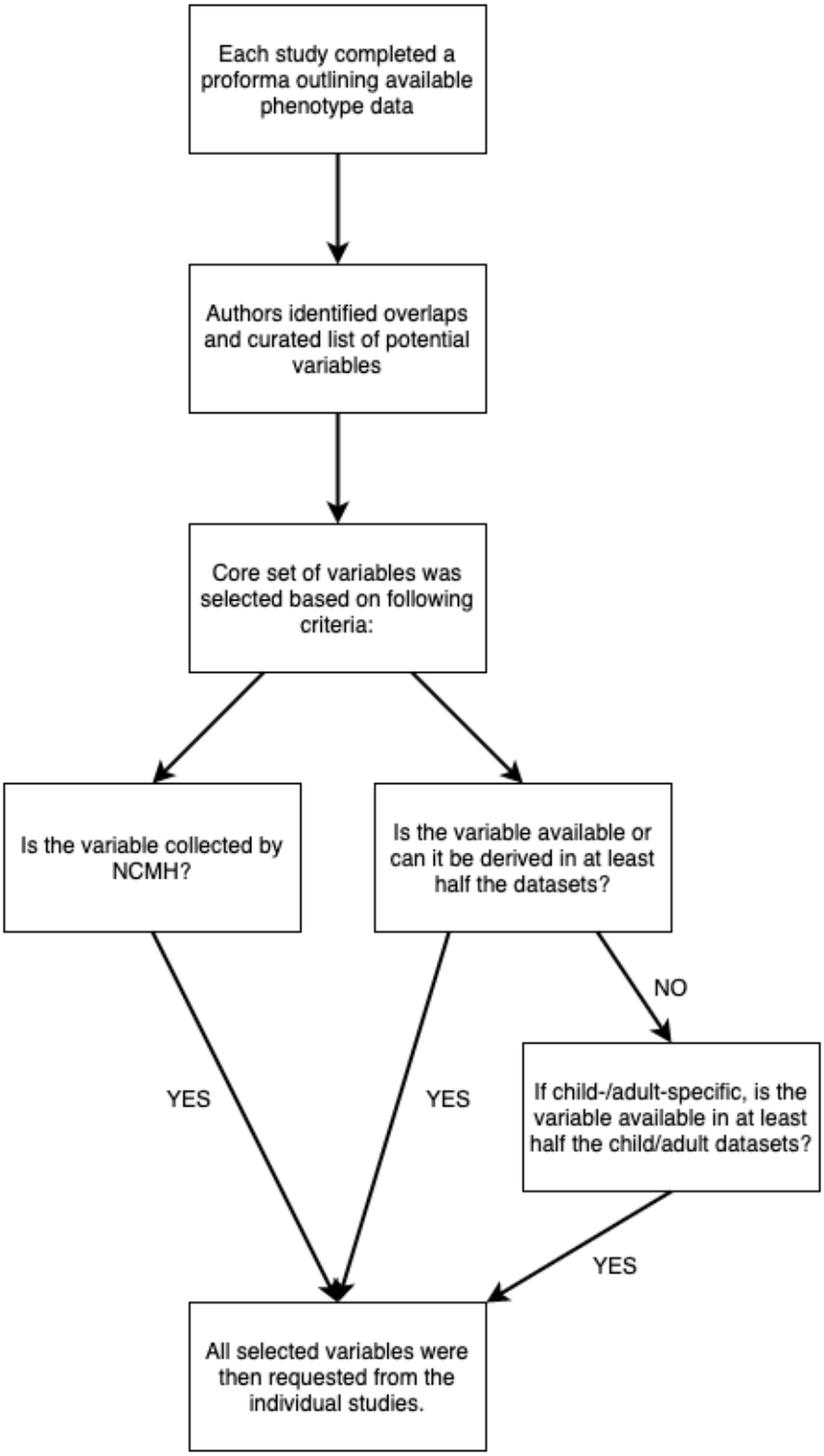


Figure 2

DRAGON-Data pipeline for SNP genotype QC and imputation

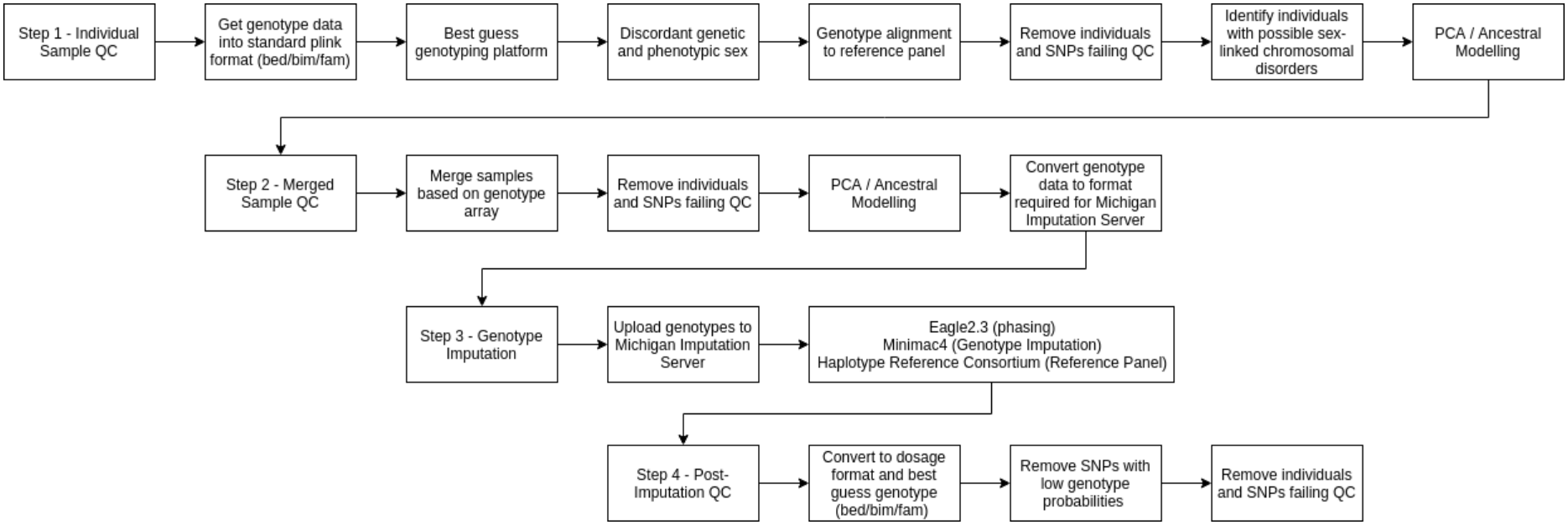


Figure 3 - **DRAGON-Data pipeline for CNV Calling**

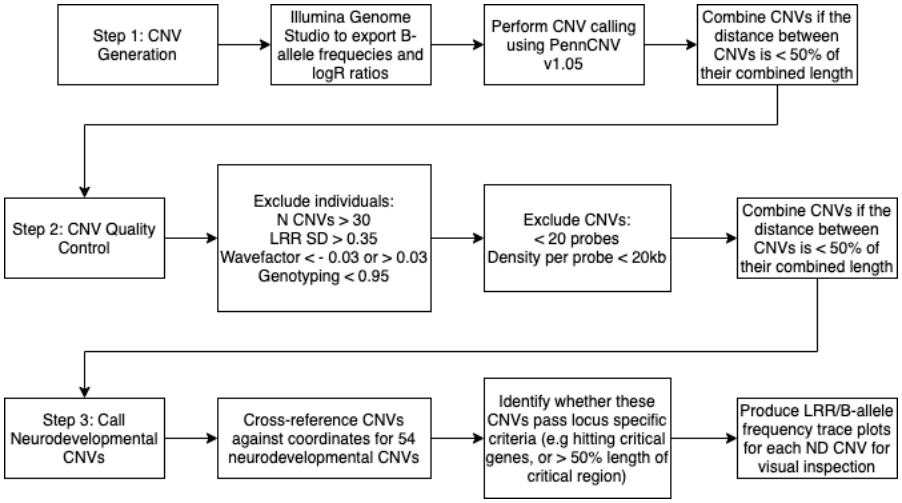
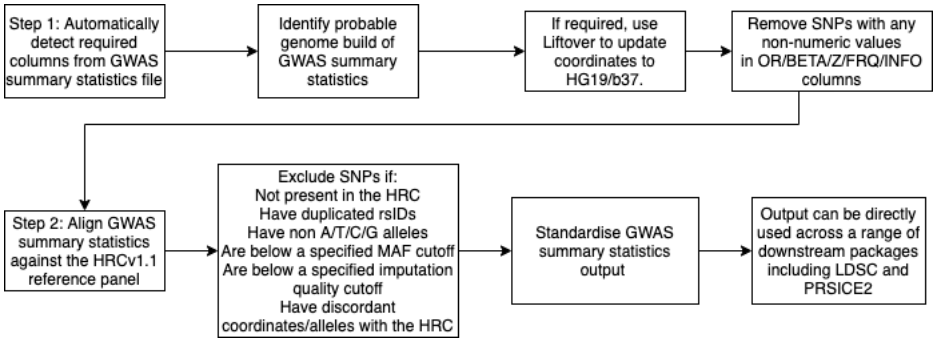


Figure 4 - DRAGON-Data Pipeline for standardising genome-wide association summary statistics (summaRygwasqc)



References

1. Atkinson P, Batchelor C, Parsons E. Trajectories of Collaboration and Competition in a Medical Discovery. *Science, Technology, & Human Values* 1998;23(3):259-84. doi: 10.1177/016224399802300301

2. Adams J. The fourth age of research. *Nature* 2013;497:557. doi: 10.1038/497557a

3. Peltonen L, McKusick VA. Dissecting Human Disease in the Postgenomic Era. *Science* 2001;291(5507):1224-29. doi: 10.1126/science.291.5507.1224

4. Iadarola ND, Niciu MJ, Richards EM, et al. Ketamine and other N-methyl-D-aspartate receptor antagonists in the treatment of depression: a perspective review. *Therapeutic Advances in Chronic Disease* 2015;6(3):97-114. doi: 10.1177/2040622315579059

5. Childress AC. A critical appraisal of atomoxetine in the management of ADHD. *Therapeutics and clinical risk management* 2015;12:27-39. doi: 10.2147/TCRM.S59270

6. MacEwan JP, Seabury S, Aigbogun MS, et al. Pharmaceutical Innovation in the Treatment of Schizophrenia and Mental Disorders Compared with Other Diseases. *Innov Clin Neurosci* 2016;13(7-8):17-25.

7. O'Donnell P, Rosen L, Alexander R, et al. Strategies to Address Challenges in Neuroscience Drug Discovery and Development. *International Journal of Neuropsychopharmacology* 2019;22(7):445-48. doi: 10.1093/ijnp/pyz027

8. Owen Michael J. New Approaches to Psychiatric Diagnostic Classification. *Neuron* 2014;84(3):564-71. doi: https://doi.org/10.1016/j.neuron.2014.10.028

9. McArthur RA. Aligning physiology with psychology: Translational neuroscience in neuropsychiatric drug discovery. *Neurosci Biobehav Rev* 2017;76:4-21. doi: https://doi.org/10.1016/j.neubiorev.2017.02.004

10. Willsey AJ, Morris MT, Wang S, et al. The Psychiatric Cell Map Initiative: A Convergent Systems Biological Approach to Illuminating Key Molecular Pathways in Neuropsychiatric Disorders. *Cell* 2018;174(3):505-20. doi: https://doi.org/10.1016/j.cell.2018.06.016

11. Smoller JW, Andreassen OA, Edenberg HJ, et al. Psychiatric genetics and the structure of psychopathology. *Mol Psychiatry* 2019;24(3):409-20. doi: 10.1038/s41380-017-0010-4

12. Denny JC, Van Driest SL, Wei W-Q, et al. The Influence of Big (Clinical) Data and Genomics on Precision Medicine and Drug Development. *Clin Pharmacol Ther* 2018;103(3):409-18. doi: 10.1002/cpt.951

13. Li J, Cai T, Jiang Y, et al. Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Mol Psychiatry* 2015;21:290. doi: 10.1038/mp.2015.40

14. Anttila V, Bulik-Sullivan B, Finucane HK, et al. Analysis of shared heritability in common disorders of the brain. *Science* 2018;360(6395):eaap8757. doi: 10.1126/science.aap8757

15. Baselmans BML, Yengo L, van Rheenen W, et al. Risk in Relatives, Heritability, SNP-Based Heritability, and Genetic Correlations in Psychiatric Disorders: A Review. *Biol Psychiatry* 2021;89(1):11-19. doi: 10.1016/j.biopsych.2020.05.034

16. Plana-Ripoll O, Pedersen CB, Holtz Y, et al. Exploring Comorbidity Within Mental Disorders Among a Danish National Population. *JAMA Psychiatry* 2019;76(3):259-70. doi: 10.1001/jamapsychiatry.2018.3658

17. van Rheenen W, Peyrot WJ, Schork AJ, et al. Genetic correlations of polygenic disease traits: from theory to practice. *Nature Rev Genet* 2019;20(10):567-81. doi: 10.1038/s41576-019-0137-z

18. Merikangas KR, Merikangas AK. Harnessing Progress in Psychiatric Genetics to Advance Population Mental Health. *Am J Public Health* 2019;109(S3):S171-S75. doi: 10.2105/AJPH.2019.304948

19. McCoy TH, Jr., Yu S, Hart KL, et al. High Throughput Phenotyping for Dimensional Psychopathology in Electronic Health Records. *Biol Psychiatry* 2018;83(12):997-1004. doi: 10.1016/j.biopsych.2018.01.011 [published Online First: 2018/03/03]

20. Sanchez-Roige S, Palmer AA. Emerging phenotyping strategies will advance our understanding of psychiatric genetics. *Nat Neurosci* 2020;23(4):475-80. doi: 10.1038/s41593-020-0609-7

21. Underwood JF, DelPozo-Banos M, Frizzati A, et al. Evidence of increasing recorded diagnosis of autism spectrum disorders in Wales, UK: An e-cohort study. *Autism*;[in press] doi: 10.1177/13623613211059674

22. Zheutlin AB, Dennis J, Karlsson Linnér R, et al. Penetrance and Pleiotropy of Polygenic Risk Scores for Schizophrenia in 106,160 Patients Across Four Health Care Systems. *The American journal of psychiatry* 2019;176(10):846-55. doi: 10.1176/appi.ajp.2019.18091085 [published Online First: 2019/08/16]

23. Crowley JJ, Sakamoto K. Psychiatric genomics: outlook for 2015 and challenges for 2020. *Current Opinion in Behavioral Sciences* 2015;2:102-07. doi: https://doi.org/10.1016/j.cobeha.2014.12.005

24. Blokland GAM, del Re EC, Mesholam-Gately RI, et al. The Genetics of Endophenotypes of Neurofunction to Understand Schizophrenia (GENUS) consortium: A collaborative cognitive and neuroimaging genetics project. *Schizophr Res* 2018;195:306-17. doi: 10.1016/j.schres.2017.09.024

25. Docherty AR, Fonseca-Pedrero E, Debbané M, et al. Enhancing Psychosis-Spectrum Nosology Through an International Data Sharing Initiative. *Schizophr Bull* 2018;44(suppl_2):S460-S67. doi: 10.1093/schbul/sby059

26. Gur RE, Bassett AS, McDonald-McGinn DM, et al. A neurogenetic model for the study of schizophrenia spectrum disorders: the International 22q11.2 Deletion Syndrome Brain Behavior Consortium. *Mol Psychiatry* 2017;22:1664. doi: 10.1038/mp.2017.161

27. Psychosis Endophenotypes International Consortium, Wellcome Trust Case-Control Consortium 2. A Genome-wide Association Analysis of a Broad Psychosis Phenotype Identifies Three Loci for Further Investigation. *Biol Psychiatry* 2014;75(5):386-97. doi: 10.1016/j.biopsych.2013.03.033

28. Sullivan PF, Agrawal A, Bulik CM, et al. Psychiatric genomics: an update and an agenda. *Am J Psychiatry* 2017;175(1):15-27.

29. Underwood JFG, Kendall KM, Berrett J, et al. Autism spectrum disorder diagnosis in adults: phenotype and genotype findings from a clinically derived cohort. *Br J Psychiatry* 2019;215(5):647-53. doi: 10.1192/bjp.2019.30 [published Online First: 2019/02/26]

30. Bath PA, Deeg D, Poppelaars JAN. The harmonisation of longitudinal data: a case study using data from cohort studies in The Netherlands and the United Kingdom. *Ageing and Society* 2010;30(8):1419-37. doi: 10.1017/S0144686X1000070X [published Online First: 2010/09/29]

31. Curran PJ, Hussong AM. Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods* 2009;14(2):81-100. doi: 10.1037/a0015914

32. McGuffin P, Farmer A, Harvey I. A polydiagnostic application of operational criteria in studies of psychotic illness: development and reliability of the OPCRIT system. *Arch Gen Psychiatry* 1991;48(8):764-70.

33. Roberts BW, DelVecchio WF. The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychol Bull* 2000;126(1):3-25. doi: 10.1037/0033-2909.126.1.3

34. Baldwin JS, Dadds MR. Reliability and Validity of Parent and Child Versions of the Multidimensional Anxiety Scale for Children in Community Samples. *J Am Acad Child Adolesc Psychiatry* 2007;46(2):252-60. doi: https://doi.org/10.1097/01.chi.0000246065.93200.a1

35. Cha CB, Franz PJ, M. Guzmán E, et al. Annual Research Review: Suicide among youth – epidemiology, (potential) etiology, and treatment. *Journal of Child Psychology and Psychiatry* 2018;59(4):460-82. doi: 10.1111/jcpp.12831

36. Klonsky ED, May AM, Saffer BY. Suicide, Suicide Attempts, and Suicidal Ideation. *Annual Review of Clinical Psychology* 2016;12(1):307-30. doi: 10.1146/annurev-clinpsy-021815-093204

37. Lally J, Watkins R, Nash S, et al. The Representativeness of Participants With Severe Mental Illness in a Psychosocial Clinical Trial. *Frontiers in Psychiatry* 2018;9(654) doi: 10.3389/fpsyt.2018.00654

38. Kline E, Hendel V, Friedman-Yakoobian M, et al. A comparison of neurocognition and functioning in first episode psychosis populations: do research samples reflect the real

world? *Soc Psychiatry Psychiatr Epidemiol* 2019;54(3):291-301. doi: 10.1007/s00127-018-1631-x

39. Martin J, Tilling K, Hubbard L, et al. Association of Genetic Risk for Schizophrenia With Nonparticipation Over Time in a Population-Based Cohort Study. *Am J Epidemiol* 2016;183(12):1149-58. doi: 10.1093/aje/kww009

40. Whitaker C, Stevelink S, Fear N. The Use of Facebook in Recruiting Participants for Health Research Purposes: A Systematic Review. *J Med Internet Res* 2017;19(8):e290. doi: 10.2196/jmir.7071 [published Online First: 28.08.2017]

41. Batterham PJ. Recruitment of mental health survey participants using Internet advertising: content, characteristics and cost effectiveness. *International Journal of Methods in Psychiatric Research* 2014;23(2):184-91. doi: 10.1002/mpr.1421

42. Judd F, Komiti A, Jackson H. How Does Being Female Assist Help-Seeking for Mental Health Problems? *Aust N Z J Psychiatry* 2008;42(1):24-29. doi: 10.1080/00048670701732681

43. Cleynen I, Engchuan W, Hestand MS, et al. Genetic contributors to risk of schizophrenia in the presence of a 22q11.2 deletion. *Mol Psychiatry* 2020 doi: 10.1038/s41380-020-0654-3

44. Wunsch G, Russo F, Mouchart M. Do We Necessarily Need Longitudinal Data to Infer Causal Relations? *BMS: Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique* 2010(106):5-18.

45. Donker T, van Straten A, Marks I, et al. Brief self-rated screening for depression on the Internet. *J Affect Disord* 2010;122(3):253-59. doi: 10.1016/j.jad.2009.07.013

46. Booth-Kewley S, Larson GE, Miyoshi DK. Social desirability effects on computerized and paper-and-pencil questionnaires. *Computers in Human Behavior* 2007;23(1):463-77. doi: 10.1016/j.chb.2004.10.020

47. Pardiñas AF, Holmans P, Pocklington AJ, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet* 2018;50(3):381-89. doi: 10.1038/s41588-018-0059-2

48. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017;186(9):1026-34. doi: 10.1093/aje/kwx246

49. Davis K, Hotopf M. Mental health phenotyping in UK Biobank. *Prog Neurol Psychiatry* 2019;23(1):4-7.

50. McGuffin P, Katz R. The Genetics of Depression and Manic-Depressive Disorder. *Br J Psychiatry* 1989;155(3):294-304. doi: 10.1192/bjp.155.3.294 [published Online First: 2018/01/02]

51. Cai N, Revez JA, Adams MJ, et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat Genet* 2020;52(4):437-47. doi: 10.1038/s41588-020-0594-5

52. Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC medicine* 2013;11(1):126.

53. Chang CC, Chow CC, Tellier L, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;4(7)

54. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* 2007;81(3):559-75. doi: 10.1086/519795

55. Rayner W. Genotyping chips strand and build files: Wellcome Centre for Human Genetics; 2018 [updated 24/03/2018. Available from: https://www.well.ox.ac.uk/~wrayner/strand/ accessed 20/08/2019].

56. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48(10):1279-83. doi: 10.1038/ng.3643

57. Deelen P, Bonder MJ, van der Velde KJ, et al. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Research Notes* 2014;7(1):901. doi: 10.1186/1756-0500-7-901

58. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. *Briefings in Bioinformatics* 2012;14(2):144-61. doi: 10.1093/bib/bbs038

59. Igo Jr. RP, Cooke Bailey JN, Romm J, et al. Quality Control for the Illumina HumanExome BeadChip. *Current Protocols in Human Genetics* 2016;90(1):2.14.1-2.14.16. doi: 10.1002/cphg.15

60. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics* 2011;27(15):2156-58. doi: 10.1093/bioinformatics/btr330

61. Patterson NJ, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006;2(12):e190. doi: 10.1371/journal.pgen.0020190

62. Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol* 2015;39(4):276-93.

63. Legge SE, Pardiñas AF, Helthuis M, et al. A genome-wide association study in individuals of African ancestry reveals the importance of the Duffy-null genotype in the assessment of clozapine-related neutropenia. *Mol Psychiatry* 2019;24(3):328-37. doi: 10.1038/s41380-018-0335-7

64. Phillips C, Parson W, Lundsberg B, et al. Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. *Forensic Sci Int Genet* 2014;11:13-25. doi: 10.1016/j.fsigen.2014.02.012

65. Kidd KK, Speed WC, Pakstis AJ, et al. Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int Genet* 2014;10:23-32. doi: 10.1016/j.fsigen.2014.01.002

66. Li JZ, Absher DM, Tang H, et al. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* 2008;319(5866):1100-04. doi: 10.1126/science.1153717

67. Chambers JC, Abbott J, Zhang W, et al. The South Asian Genome. *PLOS ONE* 2014;9(8):e102645. doi: 10.1371/journal.pone.0102645

68. Tishkoff SA, Kidd KK. Implications of biogeography of human populations for'race'and medicine. *Nat Genet* 2004;36:S21-S27.

69. Reich D, Thangaraj K, Patterson N, et al. Reconstructing Indian population history. *Nature* 2009;461:489. doi: 10.1038/nature08365

70. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016;48(10):1284-87. doi: 10.1038/ng.3656

71. Iglesias AI, van der Lee SJ, Bonnemaijer PWM, et al. Haplotype reference consortium panel: Practical implications of imputations with large reference panels. *Hum Mutat* 2017;38(8):1025-32. doi: 10.1002/humu.23247

72. Loh P-R, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* 2016;48:1443. doi: 10.1038/ng.3679

73. Wang K, Li M, Hadley D, et al. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007;17(11):1665-74. doi: 10.1101/gr.6861907

74. Kendall KM, Rees E, Escott-Price V, et al. Cognitive Performance Among Carriers of Pathogenic Copy Number Variants: Analysis of 152,000 UK Biobank Subjects. *Biol Psychiatry* 2017;82(2):103-10. doi: https://doi.org/10.1016/j.biopsych.2016.08.014

75. Stanaway IB, Hall TO, Rosenthal EA, et al. The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genet Epidemiol* 2019;43(1):63-81. doi: 10.1002/gepi.22167

76. Pistis G, Porcu E, Vrieze SI, et al. Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur J Hum Genet* 2014;23:975. doi: 10.1038/ejhg.2014.216

77. Zuvich RL, Armstrong LL, Bielinski SJ, et al. Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. *Genet Epidemiol* 2011;35(8):887-98. doi: 10.1002/gepi.20639

78. Lyon MS, Andrews SJ, Elsworth B, et al. The variant call format provides efficient and robust storage of GWAS summary statistics. *Genome Biology* 2021;22(1):32. doi: 10.1186/s13059-020-02248-0

79. Julienne H, Lechat P, Guillemot V, et al. JASS: command line and web interface for the joint analysis of GWAS results. *NAR Genomics and Bioinformatics* 2020;2(1) doi: 10.1093/nargab/lqaa003

80. MacArthur JAL, Buniello A, Harris LW, et al. Workshop proceedings: GWAS summary statistics standards and sharing. *Cell Genomics* 2021;1(1):100004. doi: https://doi.org/10.1016/j.xgen.2021.100004

81. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* 2019;8(7) doi: 10.1093/gigascience/giz082

82. Ge T, Chen C-Y, Ni Y, et al. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications* 2019;10(1):1776. doi: 10.1038/s41467-019-09718-5

83. Bulik-Sullivan BK, Loh P-R, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015;47(3):291-95. doi: 10.1038/ng.3211

http://www.nature.com/ng/journal/v47/n3/abs/ng.3211.html#supplementary-information

84. Wing JK, Babor T, Brugha T, et al. SCAN: Schedules for Clinical Assessment in Neuropsychiatry. *Archives of General Psychiatry* 1990;47(6):589-93.

85. Angold A, Costello EJ. The child and adolescent psychiatric assessment (CAPA). *Journal of the American Academy of Child & Adolescent Psychiatry* 2000;39(1):39-48.

86. Supercomputing Wales. Supercomputing Wales / Uwchgyfrifiadura Cymru 2018 [updated 27/03/2018. Available from: www.supercomputing.wales accessed 07/06/2019].

87. Boedhoe PSW, Heymans MW, Schmaal L, et al. An Empirical Comparison of Meta- and Mega-Analysis With Data From the ENIGMA Obsessive-Compulsive Disorder Working Group. *Frontiers in neuroinformatics* 2019;12:102-02. doi: 10.3389/fninf.2018.00102

88. Wojcik GL, Graff M, Nishimura KK, et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* 2019;570(7762):514-18. doi: 10.1038/s41586-019-1310-4

89. Gorski M, Günther F, Winkler TW, et al. On the differences between mega- and meta-imputation and analysis exemplified on the genetics of age-related macular degeneration. *Genet Epidemiol* 2019;43(5):559-76. doi: 10.1002/gepi.22204

90. Pulit SL, McArdle PF, Wong Q, et al. Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. *The Lancet Neurology* 2016;15(2):174-84. doi: https://doi.org/10.1016/S1474-4422(15)00338-5

91. Yang J, Lee SH, Goddard ME, et al. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am J Hum Genet* 2011;88(1):76-82. doi: http://dx.doi.org/10.1016/j.ajhg.2010.11.011

92. Gogarten SM, Bhangale T, Conomos MP, et al. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* 2012;28(24):3329-31.

93. Privé F, Aschard H, Ziyatdinov A, et al. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* 2018;34(16):2781-87. doi: 10.1093/bioinformatics/bty185

94. Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* 2015;12:115. doi: 10.1038/nmeth.3252

95. Chen H, Wang C, Conomos MP, et al. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am J Hum Genet* 2016;98(4):653-66. doi: 10.1016/j.ajhg.2016.02.012 [published Online First: 03/24]

96. Rizvi AA, Karaesmen E, Morgan M, et al. gwasurvivr: an R package for genome-wide survival analysis. *Bioinformatics* 2018;35(11):1968-70. doi: 10.1093/bioinformatics/bty920

97. Layer RM, Kindlon N, Karczewski KJ, et al. Efficient genotype compression and analysis of large genetic-variation data sets. *Nature Methods* 2015;13:63. doi: 10.1038/nmeth.3654

98. Hernaez M, Pavlichin D, Weissman T, et al. Genomic Data Compression. *Annual Review of Biomedical Data Science* 2019;2(1):[in press]. doi: 10.1146/annurev-biodatasci-072018-021229

99. EUR-LEX. General Data Protection Regulation: Publications Office of the European Union; 2016 [updated 27/04/2016. Available from: http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32016R0679 accessed 20/08/2019].

100. Erlich Y, Williams JB, Glazer D, et al. Redefining genomic privacy: trust and empowerment. *PLoS Biol* 2014;12(11):e1001983-e83. doi: 10.1371/journal.pbio.1001983

101. Rehm HL, Page AJH, Smith L, et al. GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics* 2021;1(2):100029. doi: https://doi.org/10.1016/j.xgen.2021.100029

102. Consortium PG. PGC Data Access: Open Source Philosophy Chapel Hill, North Carolina, USA: UNC School of Medicine; 2018 [Available from: https://www.med.unc.edu/pgc/shared-methods/open-source-philosophy/ accessed 15/12/2020.

103. Health Data Research UK. DATAMIND - our Hub for Mental Health Informatics Research Development 2022 [Available from: https://www.hdruk.ac.uk/helping-with-health-data/health-data-research-hubs/datamind/ accessed 10/01/2021.

104. Gordon-Smith K, Saunders K, Geddes JR, et al. Large-scale roll out of electronic longitudinal mood-monitoring for research in affective disorders: Report from the UK bipolar disorder research network. *J Affect Disord* 2019;246:789-93. doi: https://doi.org/10.1016/j.jad.2018.12.099

105. Betcheva ET, Mushiroda T, Takahashi A, et al. Case–control association study of 59 candidate genes reveals the DRD2 SNP rs6277 (C957T) as the only susceptibility factor

for schizophrenia in the Bulgarian population. *J Hum Genet* 2009;54(2):98-107. doi: 10.1038/jhg.2008.14

106. Kirov G, Zaharieva I, Georgieva L, et al. A genome-wide association study in 574 schizophrenia trios using DNA pooling. *Mol Psychiatry* 2009;14(8):796-803. doi: 10.1038/mp.2008.33

107. Hamshere ML, Walters JTR, Smith R, et al. Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC. *Mol Psychiatry* 2013;18(6):708-12. doi: 10.1038/mp.2012.67

108. Lynham AJ, Hubbard L, Tansey KE, et al. Examining cognition across the bipolar/schizophrenia diagnostic spectrum. *Journal of psychiatry & neuroscience: JPN* 2018;43(4):245.

109. Morrison S, Chawner SJRA, van Amelsvoort TAMJ, et al. Cognitive deficits in childhood, adolescence and adulthood in 22q11.2 deletion syndrome and association with psychopathology. *Translational Psychiatry* 2020;10(1):53. doi: 10.1038/s41398-020-0736-7

110. Chawner SJRA, Owen MJ, Holmans P, et al. Genotype & phenotype associations in children with copy number variants associated with high neuropsychiatric risk in the UK (IMAGINE-ID): a case-control cohort study. *The Lancet Psychiatry* 2019;6(6):493-505. doi: 10.1016/S2215-0366(19)30123-3

111. Chawner SJRA, Doherty JL, Moss H, et al. Childhood cognitive development in 22q11.2 deletion syndrome: Case–control study. *Br J Psychiatry* 2017;211(4):223-30. doi: 10.1192/bjp.bp.116.195651 [published Online First: 2018/01/02]

112. Collishaw S, Hammerton G, Mahedy L, et al. Mental health resilience in the adolescent offspring of parents with depression: a prospective longitudinal study. *The Lancet Psychiatry* 2016;3(1):49-57. doi: 10.1016/S2215-0366(15)00358-2

113. Norton N, Williams HJ, Dwyer S, et al. No evidence for association between polymorphisms in GRM3and schizophrenia. *BMC Psychiatry* 2005;5(1):23. doi: 10.1186/1471-244X-5-23

114. Lewis CM, Ng MY, Butler AW, et al. Genome-Wide Association Study of Major Recurrent Depression in the U.K. Population. *Am J Psychiatry* 2010;167(8):949-57. doi: 10.1176/appi.ajp.2010.09091380

115. Roberts NP, Kitchiner NJ, Lewis CE, et al. Psychometric properties of the PTSD Checklist for DSM-5 in a sample of trauma exposed mental health service users. *European Journal of Psychotraumatology* 2021;12(1):1863578. doi: 10.1080/20008198.2020.1863578

116. Langley K, Martin J, Agha SS, et al. Clinical and cognitive characteristics of children with attention-deficit hyperactivity disorder, with and without copy number variants. *Br J Psychiatry* 2011;199(5):398-403. doi: 10.1192/bjp.bp.111.092130 [published Online First: 2018/01/02]

117. Williams NM, Rees MI, Holmans P, et al. A Two-Stage Genome Scan for Schizophrenia Susceptibility Genes in 196 Affected Sibling Pairs. *Hum Mol Genet* 1999;8(9):1729-39. doi: 10.1093/hmg/8.9.1729