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

Amy J. Lynham*[1], Sarah Knott*[1], Jack F. G. Underwood*[1], Leon Hubbard*[1], Sharifah S. Agha[1] Jonathan I. Bisson[1], Marianne B.M van den Bree[1], Samuel J.R.A. Chawner[1], Nicholas Craddock[1], Michael O'Donovan[1], Ian R. Jones[1], George Kirov[1], Kate Langley[1,2], Joanna Martin[1], Frances Rice[1], Neil P. Roberts[1,3], Anita Thapar[1], Richard Anney[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
2. School of Psychology, Cardiff University, Cardiff, UK
3. Directorate of Psychology and Psychological Therapies, Cardiff & Vale University Health Board, Cardiff, United Kingdom

* 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

Background:

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.

Aims:

Well-characterised, cross-disorder samples are needed to investigate this matter, but currently few exist. Our aim is to develop procedures to purposely curate and aggregate genotypic and phenotypic data in psychiatric research.

Method:

As part of the Cardiff MRC Mental Health Data Pathfinder, we have curated and harmonised phenotypic and genetic information from 15 studies to create a new data repository, DRAGON-Data. To date, DRAGON-Data includes over 45,000 individuals: adults and children with neurodevelopmental or psychiatric diagnoses, affected probands within collected families and individuals who carry a known neurodevelopmental risk copy number variant (ND-CNV).

Results:

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.

Conclusions:

DRAGON-Data combines genetic and non-genetic information and is available as a resource for research across traditional psychiatric diagnostic categories. 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]. 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[2]. 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[3] and atomoxetine for attention deficit hyperactivity disorder (ADHD)[4]. Worryingly, many major pharmaceutical companies are decreasing their research efforts and investment in this area[5]. This apparent stagnation in progress is the result of a lack of understanding of the pathogenesis of these conditions[6], hindering the identification of novel targets for drug discovery, and limiting the utility of current diagnostic categories in defining mechanistically discrete disorders. A route to address these limitations involves integrating biological data at scale and across, rather than within, diagnostic classifications[7]. Research conducted in this manner can explore the aetiological and biological commonalities between diagnoses revealed by genetic studies[8], accelerating discoveries on complex disorders and informing novel pharmacological and non-pharmacological therapeutic strategies, firmly grounded in biology[9].

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[10]. 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[11]. 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[10]. 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[12]. 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[12 13]. However, some common limitations of these include selection biases and a low representation of clinically severe disorders[13 14]. The latter can be exemplified by a recent study of schizophrenia genetic liability on 106,160 patients across four US healthcare systems, where only 522 individuals with a formal diagnosis of schizophrenia were included[15], a small figure but in line with a lifetime morbid risk of 0.7% for this disorder[16]. Such is a classic quandary in psychiatric genomics[17], 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.

Aims and objectives

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 ongoing efforts to collate phenotype data within the Psychiatric Genomics Consortium (PGC)[18], as well as previous mental health-related initiatives including the Genetics of Endophenotypes of Neurofunction to Understand Schizophrenia (GENUS) consortium[19], the International Consortium for Schizotypy Research (ICSR)[20], the International 22q11.2 Deletion Syndrome Brain Behaviour Consortium (22q11.2DS IBBC)[21], the Psychosis Endophenotypes International Consortium[22], and the Genes to Mental Health (G2MH) network[23]. However, most of these projects have typically focused on a single psychiatric disorder or group of closely related conditions, while DRAGON-Data seeks to integrate genomic and phenotype data from a range of disorders across the developmental continua.

The current paper describes the formation of DRAGON-Data through the curation and harmonisation of phenotypic and genetic information across existing cohorts. These represent a broad diversity of psychiatric diagnoses including ADHD, bipolar disorder, mood disorders, major depressive disorder, neurodevelopmental conditions, post-traumatic stress disorder (PTSD) and schizophrenia. 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

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 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).

Ethics approval

The development of DRAGON-Data was reviewed by the Cardiff University School of Medicine Ethics Committee as part of the "Clinical, phenotypic and genomic research in psychiatry" application (SMREC 19/72), approved on 05/09/19. Ethical clearances to conduct each of the DRAGON-Data studies are detailed in their parent publications.

Phenotypic data harmonisation strategy

The process of curating the phenotypic data is outlined in **Figure 1**, and a description of challenges we faced in our exercise is provided in the **Supplementary Note**. 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. 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[24]. 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 fewer than half the studies were only included if they could be derived from existing data to achieve the representation threshold. On receipt of each dataset, the variables were cleaned and matched with our defined core set of variables, and these were then signposted within our DRAGON-Data dictionary.

Genetic data harmonisation strategy

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[25 26]. Genotyping platform, when not properly recorded in study logs, was inferred by comparing chromosome and basepair positions of the genotypes on each dataset and 166 array manifests[27]. Across the datasets in DRAGON-Data, Illumina chips are by far the most common (**Table 1**). Despite the standardisation inherent to genotype datasets that is driven by platform commonalities and the PLINK format conversion, creating a harmonised multi-study dataset requires stringent study-wide and dataset-wide QC. We minutely descreibe these QC steps and the challenges they are meant to address in the **Supplementary Note**.

Results

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 2** 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[28]) for adults and the Child and Adolescent Psychiatric Assessment (CAPA[29]) 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 small subgroup of this sample (n=485/16311) 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.

Key recommendations for genotype-phenotype data harmonisation

Based on our experience developing DRAGON-Data, we suggest some recommendations for the harmonisation and analysis of clinical and genetic 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 factors outlined in the Supplementary Note (study protocol differences, use of diagnostic criteria) 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 by researchers, 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, values and corresponding labels (for categorical variables), as well as 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, biological sex, gender identity, 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.
- Data should only be shared and combined if there are suitable ethical and data sharing agreements that participants have consented to. There may be separate ethical considerations for data sharing within research settings and for linkage to other external datasets, particularly public electronic health record databases.
- 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.
- 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.

Discussion

Using DRAGON-Data

All the DRAGON-Data data have been securely stored in HAWK, a high-performance computing (HPC) cluster supported by the Supercomputing Wales infrastructure[30], 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 decide whether to grant access to the relevant data on scientific grounds. This is modelled after successful international consortia such as the PGC[18], 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[31]. 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[32]. 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[25] or GCTA[33]. 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*[34] and *bigsnpr*[35]. This allows using the approximately 1,700 tools currently offered by the Bioconductor suite[36] in a large-scale genome-wide setting, and facilitates applying complex analytic techniques such as mixed-model regression[37] and survival analysis[38]. 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[25 39]. 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[40].

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)[41], the principle of data minimisation was respected, with only limited individual-level data being requested from research groups. Furthermore, as a means of maintaining 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[42], 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[43]. 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.

Limitations

Whilst there is rich demographic and clinical data available on patient cohorts in DRAGON-Data, the data on those without mental health disorders ("controls" in experimental study designs) is comparatively smaller and less detailed. The majority of the controls in DRAGON-Data came from NCMH (N=3508) and completed a brief interview that included demographic information and screened for the presence of psychiatric disorders. Four of the remaining studies in DRAGON-Data also collected data on participants without psychiatric diagnoses, but these were recruited due to being an unaffected sibling of a proband (Sib-Pairs) or by being ND-CNV carriers (ECHO, IMAGINE, DEFINE). While these samples might not be representative of a standard control population given their ascertainment, they might still be relevant for future DRAGON-Data studies. For example, merged datasets with affected, relatives of affected, and unaffected individuals have been used for research into the additivity of risk factors for neurodevelopmental traits and in the validation of polygenic score methods[44].

All the studies in DRAGON-Data predated the publication of ICD-11, which may have implications for how findings using the data translates to current clinical practice. However, DRAGON-Data includes variables covering individual symptoms, onset and duration of illness, episodes and illness course, and this data could be used to derive diagnoses according to the most recent diagnostic criteria (ICD-11 and DSM-5). There was variation across the studies in how biological sex at birth and gender identity was measured and recorded, and many studies did not include standardised questions to probe sex at birth or gender identity. This is a common problem in historical datasets and even recent census questions on sex and gender for social science research vary across countries[45]. An advantage of DRAGON-Data is the inclusion of genetic data, meaning biological sex can be identified for most participants. In addition, the largest sample with phenotype data in DRAGON-Data, NCMH, included questions for both sex at birth and gender identity.

Finally, there is limited ancestry diversity within DRAGON-Data, as all the included samples were recruited in the UK and contained a majority of individuals with European ancestry. Therefore, findings from DRAGON-Data may not be generalisable to individuals from different populations, though some cohorts (e.g. CLOZUK) can contain as much as 20% of non-European individuals from different ancestries or admixed backgrounds[46].

Open data prospects

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[47] and taking advantage of the data-sharing frameworks supported by HDR UK, such as the DATAMIND Hub[48], 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.

Data availability

All data relevant to the study are included in the article. Data from individual studies are available from multiple repositories and open resources as described in their parent publications (**Table 1**). Code for the genomic data harmonisation pipelines is available in a Github repository (https://github.com/CardiffMRCPathfinder/).

Author contributions

MJO, JH, JTRW: conceptualised and designed the study. AJL: designed and implemented the phenotypic data curation protocol into DRAGON-Data. LH, AFP: designed and implemented the genotypic data curation protocol into DRAGON-Data. AJL, LH, SK: Reviewed and implemented governance and ethical protocols into DRAGON-Data. SSA, JIB, MBMvdB, SJRAC, NC, MCOD, IRJ, GK, KL, JM, FR, NPR, AT, RA, MJO: Contributed genetic and/or phenotypic data from individual studies into DRAGON-Data. AJL, JFGU, LH, SSA, SJRAC, KL, JM, RA, AFP: Led the genetic and/or phenotypic data curation of individual DRAGON-Data studies. AJL, SK, JFGU, LH, AFP: wrote the draft of the manuscript and incorporated the revisions by the co-authors. All authors reviewed the manuscript for intellectual content, contributed to revisions and approved the final version for publication.

Funding

This work was supported by a Medical Research Council (MRC) "Mental Health Data Pathfinder" Award (MC_PC_17212). Recruitment, genotyping data collection and curation of the included samples was supported by MRC Centre (MR/L010305/1), Program (MR/P005748/1), and Project (MR/L010305/1, MR/L011166/1, MR/L011794/1, MR/N022572/1) grants to Cardiff University; funding from the Wellcome Trust (078901, 079711, 220488/Z/20/Z); Health and Care Research Wales (514032); Jules Thorn Charitable Trust (JTA/06); Economic and Social Research Council (ES/J011657/1); Action Medical Research and Baily Thomas Charitable Fund. The National Centre for Mental Health was funded by the Welsh Government through Health and Care Research Wales. We also acknowledge the support of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via Welsh Government. JFGU was supported by a Wellcome Trust GW4-CAT Clinical Doctoral Fellowship (222849/Z/21/Z). SJRAC was supported by a Medical Research Foundation Fellowship (MRF-058-0015-F-CHAW). JM was supported by a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation (27879). AFP was supported by an Academy of Medical Sciences "Springboard" award (SBF005\1083). This research was funded in whole, or in part, by the Wellcome Trust. The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Declaration of interest

MCOD, MJO, and JTRW are investigators on a grant from Takeda Pharmaceuticals Ltd. to Cardiff University, for a project unrelated to the work presented here.

References

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

2. Chawner SJ, Mihaljevic M, Morrison S, et al. Pan-european landscape of research into neurodevelopmental copy number variants: a survey by the MINDDS consortium. *European Journal of Medical Genetics* 2020:104093.

3. 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

4. Childress AC. A critical appraisal of atomoxetine in the management of ADHD. *Ther Clin Risk Manag* 2015;12:27-39. doi: 10.2147/TCRM.S59270

5. 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.

6. 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

7. 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

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

9. 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

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

11. 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

12. 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

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

14. 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

15. 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. *Am J Psychiatry* 2019;176(10):846-55. doi: 10.1176/appi.ajp.2019.18091085 [published Online First: 2019/08/16]

16. McGrath J, Saha S, Chant D, et al. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiologic reviews* 2008;30(1):67-76.

17. 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

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

19. 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

20. 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

21. 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

22. 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

23. Sébastien Jacquemont, M.D. „ Guillaume Huguet, Ph.D. „ Marieke Klein, Ph.D. „ et al. Genes To Mental Health (G2MH): A Framework to Map the Combined Effects of Rare and Common Variants on Dimensions of Cognition and Psychopathology. *American Journal of Psychiatry* 2022;179(3):189-203. doi: 10.1176/appi.ajp.2021.21040432

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

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

26. 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

27. 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].

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

29. 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.

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

31. 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

32. 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

33. Yang J, Lee SH, Goddard ME, et al. GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics* 2011;88(1):76-82. doi: http://dx.doi.org/10.1016/j.ajhg.2010.11.011

34. 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.

35. 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

36. 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

37. 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]

38. 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

39. 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

40. 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

41. 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].

42. 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

43. 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

44. Klei L, McClain LL, Mahjani B, et al. How rare and common risk variation jointly affect liability for autism spectrum disorder. *Molecular Autism* 2021;12(1):66. doi: 10.1186/s13229-021-00466-2

45. Kennedy L, Khanna K, Simpson D, et al. Using sex and gender in survey adjustment. *arXiv preprint arXiv:200914401* 2022

46. 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. *Molecular Psychiatry* 2019;24(3):328-37. doi: 10.1038/s41380-018-0335-7

47. 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.

48. 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.

49. 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. *Journal of Affective Disorders* 2019;246:789-93. doi: https://doi.org/10.1016/j.jad.2018.12.099

50. 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. *Journal of Human Genetics* 2009;54(2):98-107. doi: 10.1038/jhg.2008.14

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

52. 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. *Molecular Psychiatry* 2013;18(6):708-12. doi: 10.1038/mp.2012.67

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

54. 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.

55. 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

56. 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

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

58. 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

59. 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

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

61. 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

62. 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. *British Journal of Psychiatry* 2011;199(5):398-403. doi: 10.1192/bjp.bp.111.092130 [published Online First: 2018/01/02]

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

Figure legends

Figure 1: DRAGON-Data pipeline for phenotypic data curation.

Figure 2: DRAGON-Data pipeline for SNP genotype QC and imputation.

Table 1

Studies included in DRAGON-Data

Study	Reference	Main Diagnosis	Principal Investigator(s)	Genotyping Platform	N Genotyped (Post-QC)	Psychiatric Instruments Used	Diagnostic Criteria Included	N Phenotyped (harmonised)
BDRN	49	Bipolar disorder	N. Craddock, I. Jones, L. Jones	Affymetrix5 OmniExpress PsychChip	4806 8035 1102	SCAN	ICD-10, DSM-IV	6000
Bulgarian Trios								
Case-control data	50	Psychosis and mood disorders	G. Kirov	OmniExpress	806	SCAN	DSM-IV	305
Family data*	51	Probands with psychosis and mood disorders and their families	G. Kirov	Affymetrix6	2119	SCAN	DSM-IV	3084
CLOZUK	52 53	Treatment-resistant schizophrenia	J. T. R. Walters, M. Owen, M O'Donovan	OmniExpress	13743	None (anonymised samples)	None (anonymised samples)	16405
Cardiff COGS	54	Schizophrenia, psychosis or bipolar disorder	J. T. R. Walters, M. Owen	OmniExpress	997	SCAN	ICD-10, DSM-IV	1301
DEFINE	55	Confirmed ND-CNV carrier	J.Hall, D.Linden, M.B.M. van den Bree, M. Owen	PsychChip	971 (Number inclusive	SCID PAS-ADD	DSM-IV	125

					of ECHO and IMAGINE)			
ECHO IMAGINE	56 57	Confirmed ND-CNV carrier	M.B.M. van den Bree, J.Hall, D.Linden, M. Owen	PsychChip		CAPA	DSM-IV	963
EPAD*	58	Major depressive disorder (at least one affected parent and their child)	F. Rice, A. Thapar	PsychChip	615	CAPA and SCAN	DSM-IV	674
F-Series*	59	Psychosis and mood disorders	M. Owen	OmniExpress	749	SCAN	ICD-10, DSM-IV	1022
DeCC/DeNt	60	Major depressive disorder	N. Craddock, L. Jones, C.Lewis, M.Owen	610 Quad	1346	SCAN	DSM-IV	1504
NCMH (National Centre for Mental Health)	24	Any developmental or mental disorder	I. Jones (and others)	PsychChip	3352	SCAN (N=465) CAPS-5 PAS-ADD	For those with SCAN interviews: ICD-10, DSM-IV, DSM-5	16311
PTSD Registry	61	PTSD	J. Bisson, N. Roberts	PsychChip	325	SCID CAPS	DSM-5	325
SAGE (Study of ADHD,	62	ADHD	A. Thapar, M. O’Donovan, M.J. Owen, K.	HumanHap550 PsychChip	2073*	CAPA	ICD-10, DSM-IV	1132

Genes and Environment)*			Langley, J. Martin					
Sib-Pairs	[63]	Schizophrenia	M. Owen	OmniExpress	918	SCAN	ICD-10, DSM-IV	918

CAPA: Child and Adolescent Psychiatric Assessment; SCAN: Schedules for Clinical Assessment in Neuropsychiatry; SCID: Structured Clinical Interview for DSM-IV, CAPS-5: Clinician Administered PTSD Scale for DSM5, PAS-ADD: The Psychiatric Assessment Schedule for Adult with Developmental Disability. *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.