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. The aggregation of data from individual research groups not only maximises the utility of individual data-sets and minimises demands on participants, but enables the joint analyses of complex data that can lead to incremental advances in elucidating disease aetiology. 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 and atomoxetine for attention-deficit hyperactivity disorder (ADHD). Worryingly, many major pharmaceutical companies are decreasing their research efforts and investment in this area. This apparent stagnation in progress is the result of a lack of understanding of the pathogenesis of these conditions, 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. Research conducted in this manner can explore the aetiological and biological commonalities between diagnoses revealed by genetic studies, accelerating discoveries on complex disorders and informing novel pharmacological and non-pharmacological therapeutic strategies, firmly grounded in biology.

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. 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. 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. 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. 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. However, some common limitations of these include selection biases and a low representation of clinically severe disorders. The latter can be exemplified by a recent study of schizophrenia genetic liability on 106 160 patients across four healthcare systems in the USA, where only 522 individuals with a formal diagnosis of schizophrenia were included, a small figure, but in line with a lifetime morbid risk of 0.7% for this disorder. Such is a classic quandary in psychiatric genomics, 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 data-sets 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), as well as previous mental health-related initiatives such as the Genetics of Endophenotypes of Neurofunction to Understand Schizophrenia (GENUS) Consortium, the International Consortium for Schizotopy Research (ICSR), the International 22q11.2 Deletion Syndrome Brain Behaviour Consortium (22q11.2DS IBBC), the Psychosis Endophenotypes International Consortium and the Genes to Mental Health (G2MH) Network. However, most of these projects have typically focused on a single psychiatric disorder or group of closely related conditions, whereas 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 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.

Method

Studies included

Fifteen studies from the MRC Centre for Neuropsychiatric Genetics and Genomics at Cardiff University were included in this project. A summary of the studies can be found in Table 1. Each study had its own approved research ethics, and ethical approval for the curation and development of DRAGON-Data was obtained from Cardiff University’s School of Medicine Research Ethics Committee (approval reference 19/72). The studies included participants who were adults with psychiatric disorders, children (defined as up to age 18 years) 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 (reference 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 Supplementary Appendix 1 available at https://doi.org/10.1192/bjo.2022.636. Initially, investigators from all studies completed a pro forma detailing the data and types of measures available, including the study clinical interviews, rating scales and self-report questionnaires. We compared all of 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. As the 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 of the current data-sets, taking into consideration that some data might be specific to child or adult cohorts. Variables that were not available in the NCMH and were present in fewer than half of the studies were only included if they could be derived from existing data to achieve the representation threshold. On receipt of each data-set, 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 pipeline to facilitate standardised procedures for all aspects of genetic analysis (Fig. 2), available at https://github.com/CardiffMRCPathfinder/GenotypeQCtoHRC. The pipeline begins with conversion of genotype data into binary PLINK format. Genotyping platform, when not properly recorded in study logs, was inferred by comparing chromosome and base-pair positions of the genotypes on each data-set and 166 array manifests. Across the data-sets in DRAGON-Data, Illumina chips are by far the most common (Table 1). Despite the standardisation inherent to genotype data-sets that is driven by platform commonalities and the PLINK format conversion, creating a harmonised multi-study data-set requires stringent study-wide and data-set-wide quality control.
<table>
<thead>
<tr>
<th>Study</th>
<th>Reference</th>
<th>Main diagnosis</th>
<th>Principal investigator(s)</th>
<th>Genotyping platform</th>
<th>Number genotyped (after quality control)</th>
<th>Psychiatric instruments used</th>
<th>Diagnostic criteria included</th>
<th>Number phenotyped (harmonised)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDRN</td>
<td>24</td>
<td>Bipolar disorder</td>
<td>N. Craddock, I. Jones, L. Jones</td>
<td>Affymetrix5</td>
<td>4806</td>
<td>SCAN</td>
<td>ICD-10, DSM-IV</td>
<td>6000</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>Psychosis and mood disorders</td>
<td>G. Kirov</td>
<td>OmniExpress</td>
<td>806</td>
<td>SCAN</td>
<td>DSM-IV</td>
<td>305</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>Probands with psychosis and mood disorders and their families</td>
<td>J. T. R. Walters, M. Owen, M. O’Donovan</td>
<td>PsychChip</td>
<td>2119</td>
<td>SCAN</td>
<td>DSM-IV</td>
<td>3084</td>
</tr>
<tr>
<td></td>
<td>27,28</td>
<td>Treatment-resistant schizophrenia</td>
<td></td>
<td>OmniExpress</td>
<td>13,743</td>
<td>None (anonymised samples)</td>
<td>None (anonymised samples)</td>
<td>16,405</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>Schizophrenia, psychosis or bipolar disorder</td>
<td>J. T. R. Walters, M. Owen</td>
<td>OmniExpress</td>
<td>997</td>
<td>SCAN</td>
<td>ICD-10, DSM-IV</td>
<td>1,301</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>Confirmed ND-CNV carrier</td>
<td>J. Hall, D. Linden, M.B.M. van den Bree, M. Owen</td>
<td>PsychChip</td>
<td>971 (number inclusive of ECHO and IMAGINE)</td>
<td>SCID, PAS-ADD, CAPA</td>
<td>DSM-IV</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td>31,32</td>
<td>Confirmed ND-CNV carrier</td>
<td>M.B.M. van den Bree, J. Hall, D. Linden, M. Owen</td>
<td>PsychChip</td>
<td>615</td>
<td>CAPA and SCAN</td>
<td>DSM-IV</td>
<td>963</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>Major depressive disorder (at least one affected parent and their child)</td>
<td>F. Rice, A. Thapar</td>
<td>PsychChip</td>
<td>615</td>
<td>SCAN</td>
<td>DSM-IV</td>
<td>674</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>Psychosis and mood disorder</td>
<td>M. Owen</td>
<td>OmniExpress</td>
<td>749</td>
<td>SCAN</td>
<td>ICD-10, DSM-IV</td>
<td>1,022</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>Major depressive disorder</td>
<td>N. Craddock, L. Jones, C. Lewis, M. Owen</td>
<td>610 Quad</td>
<td>1,346</td>
<td>SCAN</td>
<td>DSM-IV</td>
<td>1,504</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>Any developmental or mental disorder</td>
<td>I. Jones (and others)</td>
<td>PsychChip</td>
<td>3,352</td>
<td>SCAN</td>
<td>ICD-10, DSM-IV, DSM-5</td>
<td>16,311</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>PTSD</td>
<td>J. Bisson, N. Roberts</td>
<td>PsychChip</td>
<td>325</td>
<td>SCID</td>
<td>CAPS, ICD-10, DSM-5</td>
<td>325</td>
</tr>
<tr>
<td></td>
<td>38</td>
<td>ADHD</td>
<td>A. Thapar, M. O’Donovan, M.J. Owen, K. Langley, J. Martin</td>
<td>HumanHap550</td>
<td>2,073^a</td>
<td>SCID</td>
<td>CAPS, ICD-10, DSM-5</td>
<td>1,132</td>
</tr>
<tr>
<td></td>
<td>39</td>
<td>Schizophrenia</td>
<td>M. Owen</td>
<td>PsychChip</td>
<td>918</td>
<td>SCAN</td>
<td>ICD-10, DSM-IV</td>
<td>918</td>
</tr>
</tbody>
</table>

BDRN, Bipolar Disorder Research Network; SCAN, Schedules for Clinical Assessment in Neuropsychiatry; COGS, Cardiff Cognition in Schizophrenia; DEFINE, Defining Endophenotypes From Integrated Neurosciences; ND-CNV, Neurodevelopmental Copy Number Variant; SCID, Structured Clinical Interview for DSM-IV; PAS-ADD, The Psychiatric Assessment Schedule for Adult with Developmental Disability; ECHO, Experiences of Children with copy number variants; IMAGINE, Intellectual Disability and Mental Health: Assessing Genomic Impact on Neurodevelopment; CAPA, Child and Adolescent Psychiatric Assessment; EPAD, Early Prediction of Adolescent Depression; DeCC/DeNt, Depression Case Control / Depression Network; NCMH, National Centre for Mental Health; CAPS-5, Clinician Administered PTSD Scale for DSM-5; PTSD, post-traumatic stress disorder; SAGE, Study of ADHD, Genes and Environment; ADHD, attention-deficit hyperactivity disorder.

^a Includes family data and/or trios.
We minutely describe these quality control steps and the challenges they are meant to address in Supplementary Appendix 1.

**Results**

**The DRAGON-Data harmonised data-set**

Table 2 displays an overview of the variables held by each study included in the final DRAGON-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 of the studies except for CLOZUK included a semi-structured clinical diagnostic interview, most commonly the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) for adults and the Child and Adolescent Psychiatric Assessment (CAPA) 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 \( (n = 485/16\,311) \) completed more detailed interviews that included 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 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:

(a) 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.

(b) Identify measures (e.g. questionnaires and interviews) that are in common across the data-sets included. These measures may be easier to harmonise for analysis, although factors outlined in Supplementary Appendix 1 (study protocol differences, use of diagnostic criteria) should be considered to ensure comparability.

(c) 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.

(d) 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.

(e) A comprehensive data dictionary should accompany the database that incorporates the information outlined above. At a minimum, each variable should have recorded the 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

---

**Table 2 List of phenotypic variables included in DRAGON-Data**

<table>
<thead>
<tr>
<th>Variables included</th>
<th>Number of studies</th>
<th>Number of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td>12</td>
<td>15,410</td>
</tr>
<tr>
<td>Mania</td>
<td>11</td>
<td>13,906</td>
</tr>
<tr>
<td>Psychosis</td>
<td>9</td>
<td>12,072</td>
</tr>
<tr>
<td>ADHD</td>
<td>4</td>
<td>2,460</td>
</tr>
<tr>
<td>Anxiety</td>
<td>4</td>
<td>2,478</td>
</tr>
<tr>
<td>Conduct disorders</td>
<td>4</td>
<td>2,460</td>
</tr>
<tr>
<td>Autism</td>
<td>4</td>
<td>2,460</td>
</tr>
<tr>
<td>PTSD</td>
<td>1</td>
<td>325</td>
</tr>
<tr>
<td>Treatment history</td>
<td>13</td>
<td>31,164</td>
</tr>
<tr>
<td>Clinical/illness history</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at onset</td>
<td>10</td>
<td>29,023</td>
</tr>
<tr>
<td>Hospital admissions</td>
<td>7</td>
<td>26,372</td>
</tr>
<tr>
<td>Suicidal ideation</td>
<td>12</td>
<td>15,410</td>
</tr>
<tr>
<td>Adverse life events</td>
<td>6</td>
<td>9,994</td>
</tr>
<tr>
<td>Education</td>
<td>9</td>
<td>24,790</td>
</tr>
<tr>
<td>Substance use</td>
<td>13</td>
<td>29,997</td>
</tr>
<tr>
<td>Family history of psychiatric illness</td>
<td>8</td>
<td>21,473</td>
</tr>
<tr>
<td>Physical health</td>
<td>11</td>
<td>27,725</td>
</tr>
<tr>
<td>Functioning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised measure of functioning (e.g. Global Assessment Scale)</td>
<td>5</td>
<td>6260</td>
</tr>
<tr>
<td>Marital/relationship status</td>
<td>7</td>
<td>23,290</td>
</tr>
<tr>
<td>Current occupation</td>
<td>7</td>
<td>25,597</td>
</tr>
<tr>
<td>Cognitive function</td>
<td>7</td>
<td>5048</td>
</tr>
</tbody>
</table>

Number of participants refers to the number of data points available for each set of variables listed. PTSD, post-traumatic stress disorder; ADHD, attention-deficit hyperactivity disorder.
they are in common across the data-sets, 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.

(f) Include basic demographic information to evaluate the representativeness of the sample, including age range, biological sex, gender identity, and ethnicity.

(g) Data-sets do not need to be combined into a single data file. A database that houses the data-sets and allows an easy combination of selected studies and variables avoids the need for a single, large-scale data-set, and minimises the computational requirements for the querying and extraction of data.

(h) 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 data-sets, particularly public electronic health record databases.

(i) Imputation should only be performed on samples that have been genotyped on the same array backbone, or where there is substantial single nucleotide polymorphism (SNP) overlap after quality control. Furthermore, when performing quality control after imputation, removal of palindromic SNPs with high minor allele frequency (>0.4) is essential to minimising batch effects for samples genotyped on different arrays.

(j) When analysing copy number variants data across arrays, potential differences in probe density and coverage mean that it is vital that plots such as those for b-allele frequency drift, number of copy number variants called per individual and logR ratio standard deviation, are visually inspected to ensure the quality of the resulting calls.

**Discussion**

**Using DRAGON-Data**

All of the DRAGON-Data data have been securely stored in HAWK, a high-performance computing cluster supported by the Supercomputing Wales infrastructure, 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, whereas 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 allow access to the relevant data on scientific grounds. This is modelled after successful international consortia such as the PGC, 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. Although the latter is relatively straightforward, jointly analysing all samples allows for a better assessment of heterogeneity in the data and can increase statistical power. 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 because of 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. Thus, the suitability of a mega-analysis or meta-analysis approach for studies using DRAGON-Data will 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 or GCTA. 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 and bigsnpr. This allows for use of the approximately 1700 tools currently offered by the Bioconductor suite in a large-scale genome-wide setting, and facilitates applying complex analytic techniques such as mixed-model regression and survival analysis. 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. However, these are active topics of research, and initiatives such as adopting the MPEG-G ISO standard will likely allow future data harmonisation projects to seamlessly incorporate whole-genome sequences.

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. First, in line with the required safeguards for processing special category data stipulated in the EU General Data Protection Regulation (GDPR; Article 89), 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, 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 (reference 19/72). As described above, procedural safeguards were put in place to ensure secure, managed access to the data-set 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 misuse.

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 before 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 past 30 years and historical studies will not all have explicit consent on the data-sharing practices that are more commonly included today. 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
Although 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 the NCMH (n = 3508), and completed a brief interview that included demographic information and screening 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 because there were an unaffected sibling of a proband (Sib-Pairs cohort) or were ND-CNV carriers (ECHO, IMAGINE and DEFINE cohorts). Although 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 data-sets 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.

All of the studies in DRAGON-Data predated the publication of the ICD-11, which may have implications for how findings obtained using the data translate 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 data-sets, and even recent census questions on sex and gender for social science research vary across countries. 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, the NCMH, included questions for both sex at birth and gender identity.

Finally, there is limited ancestry diversity within DRAGON-Data, as all of 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, although some cohorts (e.g. CLOZUK) can contain as much as 20% of non-European individuals from different ancestries or admixed backgrounds.

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 Centre for Neuropsychiatric Genetics and Genomics at Cardiff University during the past 30 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 and taking advantage of the data-sharing frameworks supported by Health Data Research UK, such as the DATAMIND Hub, 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.

Amy J. Lynham, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, UK; Sarah Knot, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, UK; Jack F. G. Underwood, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, UK; Leon Hubbard, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, UK; Jack F. G. Underwood, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, UK; Marianne B. M. van den Bree, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, UK; Jonathan I. Blisson, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, UK; Ian R. Jones, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, UK; George Kirov, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, UK; and Jules Tharm Charitable Trust (grant number JTA/06), Economic and Social Research Council (grant number ES/D161757/1) Action Medical Research and Bally 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 via the Welsh Government. J.F.G.U. was supported by a Wellcome Trust Clinical Doctoral Fellowship (grant number 222841/22/1/2). S.J.R.A.C. was supported by a Medical Research Foundation Fellowship (number MRF-058-0015-F-CHAW). J.M.P. was supported by a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation (grant number 27879). A.F.P. was supported by an Academy of Medical Sciences “Springboard” Award (number SBF0001083). This research was funded in whole, or in part, by the Wellcome Trust. The funding organisations 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

M.C.O., M.I.O. and J.T.R.W. are investigators on a grant from Takeda Pharmaceuticals Ltd. to Cardiff University, for a project unrelated to the work presented here.

References


Supplementary material

Supplementary material is available online at https://dx.doi.org/10.1192/bjp.2022.636

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 handling pipeline is available in a GitHub repository (https://github.com/CardiffMRCPathfinder).

Author contributions

M.I.O., J.H. and J.T.R.W. conceptualised and designed the study. A.I.L. designed and implemented the phenotypic data curation protocol into DRAGON Data. L.H. and A.F.P. designed and implemented the genotypic data curation protocol into DRAGON Data. All authors reviewed the governance and ethical protocols into DRAGON Data. K.S.A.A, J.I.I.B., M.B.M.v.d.B., S.J.R.A.C., M.C.O., I.R.J., G.K., K.L., J.M., R.A., and A.F.P. led the genotypic and phenotypic data curation of individual DRAGON Data studies. All authors contributed to the manuscript and interpreted 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 (number MC_PC_17/212). Recruitment, genotyping data collection and curation of the included samples was supported by MRC Centre (grant number MR/L013085/1), Program Grant (grant number MR/R024361/1 and Project grant numbers MR/L013085/1, MR/L011749/1 and MR/N22527/2) to Cardiff University; funding from the Wellcome Trust (grant numbers 092701, 079711 and 224862/202/2) to Health and Care Research Wales (grant number S41020); and Jules Tharm Charitable Trust (grant number JTA/06), Economic and Social Research Council (grant number ES/D161757/1) Action Medical Research and Bally 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 via the Welsh Government. J.F.G.U. was supported by a Wellcome Trust Clinical Doctoral Fellowship (grant number 222841/22/1/2). S.J.R.A.C. was supported by a Medical Research Foundation Fellowship (number MRF-058-0015-F-CHAW). J.M.P. was supported by a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation (grant number 27879). A.F.P. was supported by an Academy of Medical Sciences “Springboard” Award (number SBF0001083). This research was funded in whole, or in part, by the Wellcome Trust. The funding organisations 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.
Lynham et al


