Complex Psychiatry

Editorial

Complex Psychiatry 2021;7:11–15
DOI: 10.1159/000516837

Received: March 11, 2021
Accepted: April 14, 2021
Published online: April 27, 2021

The Validity of Brief Phenotyping in Population Biobanks for Psychiatric Genome-Wide Association Studies on the Biobank Scale

Jonathan R.I. Coleman

Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

The emergence of population biobanks has been of immense value to psychiatric genetics, enabling rapid increases in sample size that have partly driven recent advances in variant discovery [1–3]. The contribution of existing biobanks such as the UK Biobank has been considerable, while emerging biobanks such as All of Us offer future promise. The successful use of biobanks in genetic studies has partly inspired the development of large, targeted cohorts for psychiatric genetics, some using the recruitment power of social media. Examples of this include the Anorexia Nervosa Genetics Initiative, the Australian Sample of Depression, and the Genetic Links to Anxiety and Depression (GLAD) study [4–6]. However, several aspects of these large-scale efforts present challenges when using the data available to capture psychiatric phenotypes. In this editorial, I will discuss some of these challenges with a particular focus on the use of brief phenotyping measures. In particular, I will discuss the effects of ascertainment biases and measure validity, as well as the potential value of biobanks in understanding the genetic relationship between disorders. Throughout the editorial, I will use examples from my ongoing work related to the definition of posttraumatic stress disorder (PTSD) in the UK Biobank and GLAD.

Biobanks provide the opportunity to assess large numbers of people using standardized pipelines, often allowing many different phenotypes to be examined in a consistent manner. However, like most cohorts, biobanks suffer from ascertainment biases to varying extents (with the notable exception of comprehensive birth cohorts such as iPSYCH) [7]. For example, the UK Biobank recruited people aged between 40 and 69 years old in 2007–2010, who lived within 25 miles of one of the 22 recruitment centers [8]. The initial phenotyping process, while carefully designed to minimize the burden on participants, was nonetheless intensive, requiring several hours of travel, questioning, and physical measurements [9]. As a result, taking part in the UK Biobank required participants to have high levels of resources, time, and motivation to take part. Only 5% of those invited took part, and they were typically healthier, wealthier, and more health conscious than the invitees on average [10]. The investment of time and capital needed to participate in the UK Biobank is likely to have prevented severely and acutely mentally unwell people from taking part. Ascertainment procedures in population biobanks may result in affected participants with mild to moderate disorder severity. This has been demonstrated in several such cohorts [11–13].

karger@karger.com
www.karger.com/cxp

Correspondence to:
Jonathan R.I. Coleman, jonathan.coleman@kcl.ac.uk

Furthermore, the scale of recruitment into biobanks can worsen known issues in genetic studies. For example, subtle effects of population stratification (where both phenotype and genotype vary as a function of geographical location) can confound analyses at the biobank scale, especially when assessing the very small effects of individual common genetic variants on psychiatric traits [14, 15].

A second limitation of using many established biobanks is that they were not principally set up to study psychiatric phenotypes. For example, the UK Biobank was initiated as a study of later-life general health and morbidity [16]. Psychiatric phenotyping has become feasible as the UK Biobank has matured [17–20]. Some such phenotyping has been achieved through medical record linkage, which has numerous complexities beyond the scope of this editorial [21, 22]. Much of the rest of this phenotyping has been achieved through brief self-report questions and questionnaires, which have to be carefully designed to minimize participant burden [17]. The validity of the resultant phenotypes, compared to those obtained through clinical interviews and guided structured questionnaires, is a point of contention. In part, this reflects the comparison – while clinical interview is often viewed as a gold standard, barriers to receiving clinical attention may mean clinical diagnoses do not capture the full picture of a disorder in the population [23, 24]. It is therefore very challenging to determine whether any phenotype strategy is ultimately valid. Nonetheless, differences between clinical diagnoses and brief phenotypes affect genetic studies; this has been clearly demonstrated in the case of depression. Cai et al. [20] recently demonstrated that brief measures used in studies from the UK Biobank appear to capture genetic effects that lack specificity to depression and are instead associated with psychiatric illness more broadly. As cohort sizes for meta-analyses of psychiatric genome-wide association studies (GWASs) grow, the primary focus of these analyses will shift from discovering associated variants to the translation of findings into biological and clinical value. An increased understanding of the overarching biology of psychiatric disorders will be invaluable to treatment development and refinement, and variants broadly associated with psychiatric illness can contribute to this. Conversely, genetics could also contribute valuably to distinguishing between different conditions; for example, determining whether an individual presenting with their first episode of depression will develop major depressive disorder (MDD) or bipolar disorder, which have different primary treatment strategies [25]. For genetics to be valuable in this way, GWASs must be able to detect distinct associations between different phenotypes. Phenotypes that capture general psychiatric illness will not contribute usefully to such efforts.

Brief measures used to assess psychiatric disorders in biobanks may not capture the same disorders as observed in the clinic, with implications for genetic analyses. Furthermore, these measures may not be appropriate for assessing these phenotypes in the general population. Many people with common mental health symptoms of clinical concern do not seek medical help for their symptoms [23, 24]. As such, extending psychiatric assessment beyond help-seeking individuals would be valuable, and widespread assessment in population biobanks offers a means to achieve this. However, many of the brief phenotypic measures deployed in biobank research were not designed for broad population research, but for assessing help-seeking individuals. Taking the example of PTSD, the 6-item shortened PTSD Checklist (PCL-C) questionnaire was coopted for use in the UK Biobank. It was initially designed for assessing PTSD symptoms in a primary care setting, where it demonstrated good psychometric properties [26]. Its use in the UK Biobank is a departure from its original intended use. Furthermore, in cohorts assessed for PTSD, there is typically a clear focal traumatic event to which the PCL-C assesses psychopathological responses. In the UK Biobank, this focus on a specific trauma is lost. In this context, there is a risk that these questions may be capturing general common psychopathological response to trauma more broadly, rather than specific PTSD pathology in the context of a focal trauma. This would then have knock-on effects on the specificity of genetic associations with the phenotype, akin to those noted by Cai et al. [20] in the case of depression.

We can already assess, to an extent, whether this last concern about nonspecificity is realized. In the most recent GWAS from the Psychiatric Genomics Consortium (PGC) PTSD working group, PTSD from the UK Biobank was defined using cutoffs on the PCL-C [3]. GWAS data from the UK Biobank were then contrasted with GWAS data from PTSD phenotypes obtained from studies using clinical interview or guided structured questionnaires from the PGC. Some differences were apparent; for example, the heritability of PTSD in men was higher in the UK Biobank than in the PGC. However, the genetic correlation between the 2 sets of GWAS data was high (0.73), and the 2 GWASs were deemed sufficiently similar that meta-analyzing them would be informative for understanding the genetics of PTSD [3]. In further research,

we have recently explored how the genetics of MDD in the presence of trauma exposure compares to the genetics of PTSD, using PTSD data both from population biobanks (UK Biobank) and from clinically ascertained samples (PGC and the Million Veteran Program). We found tentative evidence that PTSD and MDD with trauma exposure are genetically similar, but separable [27]. However, our results were similar when comparing PTSD diagnoses from population biobanks and from clinically ascertained samples. As such, we provided some evidence for a shared genetic component underlying a psychopathological response to trauma, but our results do not strongly support the idea that brief PTSD phenotyping in the UK Biobank reflects this component more than does PTSD phenotyping in clinically ascertained samples. Nonetheless, the impact of brief phenotyping in PTSD research is still unresolved – there is good evidence of a substantial shared genetic component with clinically ascertained studies, but sufficient differences that further research is needed. This is compounded by apparent sex differences and differences in the nature of trauma exposure, meaning that careful, stratified studies are required to understand the genetic heterogeneity of PTSD [3, 28].

One benefit of the emergence of biobanks is that it gives us the possibility of understanding and adapting to these challenges. An example of this can be seen in the GLAD study. The primary purpose of the GLAD study was to establish a recontactable research resource for studying MDD, anxiety disorders, and related phenotypes, including PTSD [4]. Sample recruitment was primarily driven by online advertising through social media, overcoming some of the access issues that created biases in the UK Biobank (although this recruitment strategy creates its own ascertainment biases as well). The baseline phenotyping of the GLAD study included the PCL-C phenotyping seen in the UK Biobank, but the recontactable nature of the resource has allowed us to also ask participants to volunteer to complete the full PTSD Checklist, providing us with more in-depth data on these participants. Furthermore, we can also recontact participants for clinical interviews in the future. Traditional challenges of genetic epidemiology still apply, including the difficulties of getting in-depth phenotyping on sufficiently large numbers of participants to enable well-powered analyses. Nonetheless, we have the potential to undertake detailed psychometrics on the brief measures used in biobanks and to establish key construct validity scores for future large-scale research in cohorts containing nonhelp-seeking individuals. By mirroring the brief phenotyping strategy of the UK Biobank in smaller cohort samples – amenable to targeted reanalysis using more detailed approaches – we can undertake sensitivity studies that directly inform the wider use of brief measures in psychiatric research.

Psychiatric traits are highly comorbid and show considerable genetic pleiotropy. It is therefore an open question whether genetic influences on conditions such as PTSD and MDD can be meaningfully distinguished, even without considering the validity of specific phenotyping approaches. Previous research has hypothesized a single latent variable, referred to as the p-factor, onto which different psychiatric traits load to varying degrees [29–31]. This implies that many genetic variants associated with a given psychiatric trait may not be specific to that trait, but may instead be associated with psychopathology in general. Recent work extends this idea, suggesting that a more complex model might be more appropriate [32]. Specifically, this work uses genetic correlations between GWASs of 11 psychiatric disorders to propose a loose 4-factor solution, separating compulsive, psychotic, neurodevelopmental/hyperarousal, and internalizing disorders [32]. In this model, PTSD loads primarily on the neurodevelopmental/hyperarousal factor (driven by a strong genetic correlation with ADHD), but also on the internalizing factor. These genetic similarities could argue for grouping disorders into broad clusters; for example, examining MDD, anxiety disorders, and PTSD as "internalizing disorders" [33]. In counterpoint to this, there are some clear phenotypic separations between the disorders, notably the symptom of reexperiencing in PTSD, which does not have a clear analog in anxiety disorders or MDD. This argues instead for adopting a symptom-wise approach. This is another strength of detailed biobanking. Conducting symptom-level analyses is easier when using consistent symptom measurements from a large biobank than when using aggregated, disparate clinical studies from consortia. Multivariate methods for GWAS built around genetic correlations [32] offer promise for distinguishing the shared and specific genetic influences on psychiatric symptoms. This has already been demonstrated for genetic influences on distinct symptoms of PTSD [34]. Using data from 186,689 participants of the Million Veteran Program, shared and distinct genetic associations with the reexperiencing, avoidance, and hyperarousal symptoms of PTSD were observed, as well as genetic associations with overall PTSD symptoms and diagnosis. Furthermore, PTSD symptoms were separable from internalizing traits and disorders (anxiety disorders, depression, and neuroticism), and what shared genetic component was present appeared to act via hy-

perarousal symptoms, rather than via reexperiencing or avoidance [34]. Biobanks therefore offer the potential for robust investigation of psychiatric illness at the symptom level.

In summary, population biobanks are of clear value in genetic research, providing large numbers of consistently assessed participants. Nonetheless, the phenotypes and results obtained in these biobanks require careful interpretation and understanding by researchers, especially given that it is reasonable to expect that important biases will differ between biobanks and between phenotypes. It is vital that researchers using data from biobanks understand the particular biases underlying their results and contrast their results with data obtained from other sources, ideally including studies using more intensive, traditional measures. This can provide a triangulation of evidence and reinforce conclusions against overreliance on potentially biased evidence from a single study [35].

It is also important that researchers communicate the biases affecting their work. A paradoxical strength of the UK Biobank is that the ascertainment biases it exhibits have been widely and openly discussed [8–10, 14–16], and so are much better appreciated than those of other cohorts. Cohort profiles of emerging biobanks and large cohorts should similarly promote their potential biases. Researchers working on biobank data should explicitly comment on how such biases affected their work and should justify the many analytical decisions that are made in analyzing such data. Efforts should also be made to establish large, recontactable cohorts, with parallel brief and in-depth phenotyping. These can contribute not just from the research they produce, but also from the context they give to research from biobanks.

The question is not whether results from population-level biobanks are useful – the homogeneity and scale of population biobank research are evident strengths. Instead, the challenge to researchers is to understand the nature of the specific phenotypes they study through careful research that makes use of those strengths. Population biobanks present an important source of data for exploring genetic effects at a symptom level. Integrating these data with those from clinical studies and large recontactable cohorts will yield new insights into psychiatric traits and disorders.

Acknowledgment

The author is grateful to Kirstin Purves, Jessica Mundy, and Laura Blackie for their comments on an earlier draft of this manuscript.

Conflict of Interest Statement

J.R.I.C. is an editorial board member for Complex Psychiatry but has no other conflicts of interest to declare.

Funding Sources

No funding was needed for this study.

Author Contributions

J.R.I.C. conceived and wrote this manuscript.

References

1 Mullins N, Forstner AJ, O'Connell KS, Coombes B, Coleman JRI, Qiao Z, et al. Genome-wide association study of over 40,000 bipolar disorder cases provides novel biological insights. Nat Genet. 2021 May 17. https://doi.org/10.1038/s41588-021-00857-4.

2 Howard DM, Adams MJ, Clarke TK, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci. 2019 Feb 4;22:343–52.

3 Nievergelt CM, Maihofer AX, Klengel T, Atkinson EG, Chen CY, Choi KW, et al. International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci. Nat Commun. 2019 Oct 8;10:4558.

4 Davies MR, Kalsi G, Armour C, Jones IR, McIntosh AM, Smith DJ, et al. The Genetic Links to Anxiety and Depression (GLAD) Study: online recruitment into the largest recontactable study of depression and anxiety. Behav Res Ther. 2019 Oct 24;123:103503.

5 Thornton LM, Munn-Chernoff MA, Baker JH, Juréus A, Parker R, Henders AK, et al. The Anorexia Nervosa Genetics Initiative (ANGI): overview and methods. Contemp Clin Trials. 2018 Oct 1;74:61–9.

6 Byrne EM, Kirk KM, Medland SE, McGrath JJ, Colodro-Conde L, Parker R, et al. Cohort profile: the Australian genetics of depression study. BMJ Open. 2020 May 26;10:e032580.

7 Schork AJ, Won H, Appadurai V, Nudel R, Gandal M, Delaneau O, et al. A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. Nat Neurosci. 2019 Jan 28;22:353–61.

8 Allen NE, Sudlow C, Peakman T, Collins R; UK Biobank. UK biobank data: come and get it. Sci Transl Med. 2014 Feb 19;6:224ed4.

9 Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015 Mar 31;12:e1001779.

10 Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. Am J Epidemiol. 2017 Nov 1;186:1026–34. https://doi.org/10.1093/aje/kwx246.

11 Davis KAS, Cullen B, Adams M, Brailean A, Breen G, Coleman JRI, et al. Indicators of mental disorders in UK Biobank: a comparison of approaches. Int J Methods Psychiatr Res. 2019 Aug 8;28:e1796.
12 Knudsen AK, Hotopf M, Skogen JC, Overland S, Mykletun A. The health status of nonparticipants in a population-based health study: the Hordaland Health Study. Am J Epidemiol. 2010 Dec 1;172:1306–14.
13 Taylor AE, Jones HJ, Sallis H, Euesden J, Stergiakouli E, Davies NM, et al. Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. Int J Epidemiol. 2018 Aug 1;47:1207–16.
14 Abdellaoui A, Verweij KJH, Nivard MG. Geographic confounding in genome-wide association studies. BioRxiv. 2021 Mar 18. https://doi.org/10.1101/2021.03.18.435971.
15 Haworth S, Mitchell R, Corbin L, Wade KH, Dudding T, Budu-Aggrey A, et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. Nat Commun. 2019 Jan 18;10:333.
16 Ollier W, Sprosen T, Peakman T; UK Biobank. UK Biobank: from concept to reality. Pharmacogenomics. 2005 Sep;6:639–46.
17 Davis KAS, Coleman JRI, Adams M, Allen N, Breen G, Cullen B, et al. Mental health in UK Biobank – development, implementation and results from an online questionnaire completed by 157 366 participants: a reanalysis. BJPsych Open. 2020 Feb 6;6:e18.
18 Smith DJ, Nicholl BI, Cullen B, Martin D, Ul-Haq Z, Evans J, et al. Prevalence and characteristics of probable major depression and bipolar disorder within UK biobank: cross-sectional study of 172,751 participants. PLoS One. 2013 Nov 25;8:e75362.
19 Fabbri C, Hagenaars SP, John C, Williams AT, Shrine N, Moles L, et al. Genetic and clinical characteristics of treatment-resistant depression using primary care records in two UK cohorts. Mol Psychiatry. 2021 Mar 22. https://doi.org/10.1038/s41380-021-01062-9.
20 Cai N, Revez JA, Adams MJ, Andlauer TFM, Breen G, Byrne EM, et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. Nat Genet. 2020 Mar 30;52:437–47.
21 Smoller JW. The use of electronic health records for psychiatric phenotyping and genomics. Am J Med Genet B Neuropsychiatr Genet. 2018;177:601–12.
22 Beesley LJ, Salvatore M, Fritsche LG, Pandit A, Rao A, Brummett C, et al. The emerging landscape of health research based on biobanks linked to electronic health records: existing resources, statistical challenges, and potential opportunities. Stat Med. 2020 Mar 15;39:773–800.
23 McManus S, Bebbington P, Jenkins R, Brugha T. Mental health and wellbeing in England: Adult Psychiatric Morbidity Survey 2014: a survey carried out for NHS digital by NatCen Social Research and the Department of Health Sciences, University of Leicester. Leeds: NHS Digital; 2016.
24 Henderson C, Evans-Lacko S, Thornicroft G. Mental illness stigma, help seeking, and public health programs. Am J Public Health. 2013 May;103:777–80.
25 Liebers DT, Pirooznia M, Ganna A, Goes FS, Bipolar Genome Study (BiGS). Discriminating bipolar depression from major depressive disorder with polygenic risk scores. Psychol Med. 2020 Feb 17;1–8. http://dx.doi.org/https://doi.org/10.1017/s003329172000015x.
26 Lang AJ, Stein MB. An abbreviated PTSD checklist for use as a screening instrument in primary care. Behav Res Ther. 2005 May;43:585–94.
27 Mundy J, Huebel C, Gelernter J, Levey DF, Murray RM, Skelton M, et al. Psychological trauma and the genetic overlap between posttraumatic stress disorder and major depressive disorder. medRxiv. 2020 Nov 27. https://doi.org/10.1101/2020.11.25.20229757.
28 Huckins LM, Chatzinakos C, Breen MS, Hartmann J, Klengel T, da Silva Almeida AC, et al. Analysis of genetically regulated gene expression identifies a prefrontal PTSD gene, SNRNP35, specific to military cohorts. Cell Rep. 2020 Jun 2;31:107716.
29 Caspi A, Houts RM, Belsky DW, Goldman-Mellor SJ, Harrington H, Israel S, et al. The p factor: one general psychopathology factor in the structure of psychiatric disorders? Clin Psychol Sci. 2014 Mar;2:119–37.
30 Allegrini AG, Cheesman R, Rimfeld K, Selzam S, Pingault JB, Eley TC, et al. The p factor: genetic analyses support a general dimension of psychopathology in childhood and adolescence. J Child Psychol Psychiatry. 2020;61:30–9.
31 Sprooten E, Franke B, Greven CU. The P-factor and its genomic and neural equivalents: an integrated perspective. Mol Psychiatry. 2021 Feb 1.
32 Grotzinger AD, Mallard TT, Akingbuwa WA, Ip HF, Adams MJ, Lewis CM, et al. Genetic architecture of 11 major psychiatric disorders at biobehavioral, functional genomic, and molecular genetic levels of analysis. medRxiv. 2020 Sep 23. https://doi.org/10.1101/2020.09.22.20196089.
33 Kotov R, Krueger RF, Watson D, Achenbach TM, Althoff RR, Bagby RM, et al. The Hierarchical Taxonomy of Psychopathology (HiTOP): a dimensional alternative to traditional nosologies. J Abnorm Psychol. 2017 May;126:454–77.
34 Stein MB, Levey DF, Cheng Z, Wendt FR, Harrington K, Pathak GA, et al. Genome-wide association analyses of post-traumatic stress disorder and its symptom subdomains in the Million Veteran Program. Nat Genet. 2021 Feb;53(2):174–84.
35 Taylor AE, Munafò MR. Triangulating meta-analyses: the example of the serotonin transporter gene, stressful life events and major depression. BMC Psychol. 2016 May 31;4:23.