Schizophrenia Bulletin vol. 39 no. 5 pp. 955–959, 2013
doi:10.1093/schbul/sbt108
Advance Access publication August 1, 2013

Where GWAS and Epidemiology Meet: Opportunities for the Simultaneous Study of Genetic and Environmental Risk Factors in Schizophrenia

John J. McGrath*[,1,2], Preben Bo Mortensen[3–5], Peter M. Visscher[1], and Naomi R. Wray[1]

[1]Queensland Brain Institute, University of Queensland, St Lucia, Queensland, Australia; [2]Queensland Centre for Mental Health Research, The Park Centre for Mental Health, Richlands, Australia; [3]National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark; [4]Centre for Integrated Register-based Research, CIRRAU, Aarhus University, Aarhus, Denmark; [5]The Lundbeck Foundation Initiative for Integrative Psychiatric Research, The Initiative for Integrative Psychiatric Research, Aarhus University, Aarhus, Denmark

*To whom correspondence should be addressed; Queensland Brain Institute, University of Queensland, St Lucia, Queensland 4072, Australia; tel: +61-7-3271-8694, fax: +61-7-3271-8698, e-mail: j.mcgrath@uq.edu.au

Key words: genome-wide association study/epidemiology/Mendelian randomization/interaction/polygene score

Introduction

Epidemiologists and geneticists have tended to explore their respective domains independently, and as a consequence, these fields have drifted apart. The aim of this article is to outline how clues from genetics and epidemiology can be combined in order to optimize schizophrenia research. In particular, we will explore how recent developments in statistical genetics might be used to leverage clues from epidemiology. The ultimate goal of this type of research is to help generate better treatments for schizophrenia via the identification of shared pathways[1] and to find modifiable risk factors that could lead to the primary prevention of schizophrenia.[2]

Epidemiological Frameworks Can Improve Sampling for Genetic Studies

Epidemiological studies in schizophrenia have been based on population-based surveys, case-control studies, or cohort studies. Access to large, population-based mental health registers, especially in Nordic countries, has strongly influenced psychiatric epidemiology. These studies have provided population-based estimates for a wide range of risk factors (eg, urban birth, paternal age, psychiatric family history, and infections) and have been strengthened by design features such as the avoidance of selection bias and control of multiple confounders. In contrast, genetic samples have traditionally been recruited from (*a*) informative samples (eg, multiplex families and isolated populations) or (*b*) convenient samples (eg, cases recruited from attendees at mental health services). The recruitment of large, representative samples of cases from national registers is feasible in some nations. While some research precision is lost through the use of administrative diagnoses, this can be balanced against the gain in accuracy through large sample sizes and the knowledge that subsequent results are broadly representative of national clinical practice. The recruitment of large, representative samples of controls is more widely achievable. For example, The Wellcome Trust Case Control Consortium used the large UK 1958 birth cohort as the "healthy control" sample in their landmark studies.[3] Studies related to the genetics of schizophrenia that have recruited large samples based on national registers have been published.[4–6]

Combining Clues From Genetics and Epidemiology—Measures Related to Family History and Heritability Estimates

In the absence of direct genetic data, researchers have relied on various parameters derived from familial aggregation. These have ranged from simple dichotomies (eg, family history present vs absent) to more quantitative metrics (eg, sibling recurrence risk ratio and heritability estimated from cohorts of twins or families). The recurrence of a disorder in different family members suggests that genetic factors and/or shared environmental factors within the family contribute to disease risk. By examining the risk of the disorder in certain relatives (eg, first-degree vs second-degree relatives and monozygotic vs dizygotic twins) and by making assumptions about the degree of shared environment (eg, twins raised together are more likely to experience similar environments), attempts can be made to separate contributions of genetic and non-genetic factors providing clues to the risk architecture of disorders. Parameters such as heritability[7] and sibling recurrence risk[8] have played an important role in guiding genetic studies over recent decades.

Patients with schizophrenia are routinely asked about the presence or absence of other family members with this disorder—indeed family history can be a feature in some diagnostic checklists. While easy to ask, this variable has its limitations as a measure of genetic susceptibility because the absence of known family history may be a poor measure of heritable factors[9] and family history may be poorly recorded.[10] Furthermore, a family history of a broad range of nonpsychotic disorders has also been linked to the risk of schizophrenia.[11] Despite these limitations, family history is an important positive predictor[12,13] and has been informative in schizophrenia research—studies have explored the association between (*a*) urbanicity and family history,[14] (*b*) cannabis use and family history,[15,16] and (*c*) trauma exposure and family history.[17]

In twin or family studies, it is feasible to stratify subjects by candidate exposures of interest in order to explore the impact of this variable on estimates of heritability or sibling recurrence risk.[18] The relative influence of environmental factors to heritability estimates for many neuropsychiatric phenotypes varies across the life span—thus stratifying heritability estimates by age range can provide clues to the relative contribution of environmental factors.[19,20]

Gene by Environment Interactions—A Few Words of Caution

"Genotype by environment interaction" (G × E) is a term frequently used in genetic epidemiology but remains the cause of much confusion—the term is used in different contexts and with different definitions. Some researchers use G × E to mean that the outcome (in our case schizophrenia) is the result of a process that involves both genetic and environmental risk factors. Others use G × E in a strict statistical sense. The problem with interactions in general is that they depend on the model and scale. This is not a trivial matter,[21,22] as the next scenario demonstrates. First, we assume that the prevalence of schizophrenia is 1% (an absolute risk of 0.01). Next, we take a hypothetical environmental risk factor with a relative risk of 2 (absolute risk of 0.02) and a hypothetical genetic risk factor with a relative risk of 3 (absolute risk of 0.03). If the relative risk of an individual with both risk factors is 5 (absolute risk of 0.05), then some would say this is G × E because the environmental and genetic risk factors work together to increase risk. Others might say that there is no G × E because the total risk is just the sum of the two risk factors (2 + 3 = 5). A third interpretation is that there is statistical evidence for negative G × E because the total risk is less than the product of the risk factors (5 and not 2 × 3 = 6).

In epidemiology and other fields of research, a model to explain an all-or-not outcome as a function of multiple explanatory factors is usually nonlinear. Logistic regression is just one example of this. The reason for such models is that they fit empirical data better. This implies that at some unobserved scale (eg, an unobserved scale of liability), there are explanatory variables that act additively but that at the scale of observation (0–1; unaffected–affected), the effects are nonadditive (usually multiplicative). For example, Zammit and colleagues[22] investigated the effects of multiple environmental risk factors for psychosis and found that a multiplicative model on the observed scale fitted the empirical data much better than an additive model on that scale. We tend to think of hypotheses on the scale of observation, hence it is widely expected that gene by environment interactions will underpin the genetic architecture of disorders like schizophrenia even though appropriate statistical analyses are likely to be undertaken on a transformed scale where the G and E act additively. Thus, care needs to be used when moving between different scales (disease scales vs liability scales) and their associated models. Zammit and colleagues[22] have also stressed that a good model fit does not say anything about the mechanism or underlying biology.

Combining Clues From Genetics and Epidemiology—Specific Candidate Gene Studies

The evidence linking risk of psychiatric disorders vs (*a*) specific single nucleotide polymorphisms (SNPs) in particular genes and (*b*) specific environmental exposures has been disappointing. Commentators have noted that the field is characterized by lack of replication, diminishing effect size over time, and probable publication biases.[23] Studies have examined gene by environment interactions in schizophrenia. For example, van Os and colleagues have examined the links between (*a*) cannabis use and polymorphisms in genes involved in related dopaminergic pathways[24,25] and (*b*) variants in stress pathways and vulnerability to psychotic-like experience.[26]

The wisdom of examining SNPs and environmental factors one at a time has been overshadowed by the highly polygenic nature of schizophrenia.[27,28] The large sample sizes brought together through the international Psychiatric Genomics Consortium (http://pgc.unc.edu/) for schizophrenia is yielding important results.[29,30] The statistical tools are in place to study G × E in large data sets, but progress is limited by the availability of environmental risk factors recorded in a uniform manner across large and disparately collected cohorts.

Advances in technology mean that we are no longer limited by genotyping but by phenotyping. While costs of genotyping have declined, the costs of sample acquisition and characterization have not. Kohane et al[31] has argued persuasively for the use of electronic health records in order to retrieve phenotypic data. Perlis and colleagues[32] have demonstrated proof of principle through extraction of a large sample of major depressive disorder cases classified as either treatment resistant or treatment responsive from the clinical narrative notes stored as electronic health records.

Linking Genetics and the Environment—Genome-Wide Approaches

Schizophrenia research has made important methodological contributions to the field of statistical genetics. For example, SNP-derived, genome-wide metrics can capture disease-risk estimates in the absence of knowing the precise nature of the individual risk alleles.[33] These methods were first applied to the schizophrenia genome-wide association study data.[34] Briefly, the method uses odds ratios estimated from the association analysis of a "discovery" sample. Subsequently, in an independent sample (the replication set), a score can be derived (ie, a polygene profile score) based on the presence or absence of the risk allele and the effect size in the discovery set.

Using polygenic scores, compared with the use of psychiatric family history as a measure of genetic liability, may provide several distinct advantages. First, they can be applied to the whole population of cases and controls insofar one can get access to the necessary biological material. Second, it will provide a continuous measure of liability rather than the dichotomous or categorical measure of family history. These features mean that studies using these measures should have a greatly enhanced power to study the impact of G and from the perspective of mutual confounding, mediation as well as interaction. All in all, one could justifiably hope that the rapidly increasing understanding of the polygenic architecture underpinning schizophrenia liability will enhance studies exploring the environmental causes of schizophrenia. For the first time, it will be possible to derive valid continuous measures for genetic liability on an individual level.

Polygene profile scores can also make other important contributions to research related to the combining of genetic and environmental information. For example, the addition of information on environmental exposures may increase the predictive value of a polygene profile score in independent case-control samples. In this way, the information from a polygene profile score generated from large genetically informative cohorts can then be used in much smaller cohorts that have data on both genetic and environmental risk factors. In case-only studies, exploring the correlation between polygene profile scores vs continuous measures of exposures of interest (eg, cannabis use, neonatal vitamin D concentration) may provide clues to guide future research. When sampling for epidemiological studies, it may be feasible to stratify samples by polygene profile score strata in order to explore the underlying risk architecture. It is feasible that such stratification may, in certain circumstances, improve the power of the analysis, eg, by stratifying cases who are "more genetic" (higher polygenic risk score) vs those who are more "environmental" (lower polygenic risk score). It may also be feasible to generate polygene profile subscores that are conditioned on prior hypotheses (eg, scores based on subsets of genes in particular biological pathways of interest or genes related to environmental exposures of interest).

More recently, methods have been developed that use genome-wide similarities in genotypes between all pairs of individuals have been able to estimate the proportion of variance tagged by SNPs for quantitative traits[35] and for case-control traits.[36] As with the polygene scores described above, if the cases could be stratified according to exposures of interest, then genome-wide similarities may be greater within these strata vs between the strata.

Mendelian Randomization and SNP-Based Instruments Related to Environmental Exposures

A special form of polygenic risk prediction is polygenic Mendelian Randomization (MR). Whereas the motivation of polygenic risk prediction is association or correlation, the motivation of MR is causality. MR uses genotype as an "instrumental variable" (IV). Briefly, MR[37,38] is a method that uses genetic information to quantify and test the causal effect of an exposure variable on disease or other outcomes in studies that are not experimental/randomized by design, even in the presence of confounding factors. The idea is that an unbiased effect of a modifiable exposure variable can be estimated without doing a randomized trial, by using the properties of genetic polymorphisms in the population that have an effect on the exposure variable. The assumption is that a genotype affects disease indirectly through its direct effect on the exposure variable. Under the assumption of random mating, genotypes in the population are randomized with respect to confounders. These confounders may lead to biased estimates of putative "causal" effects in observational epidemiological studies. A classic application of MR demonstrated that although C-reactive protein (CRP) levels are elevated in those with metabolic syndrome, CRP is not causal of metabolic syndrome (in an analysis in which CRP haplotypes were the IV).[39] More recently, it has been demonstrated that although smoking rates are high in those with anxiety and depression, smoking does not appear to be causal of anxiety and depression, based on an analysis in which SNP rs1051730 (robustly associated with smoking quantity and nicotine dependence) was the IV.[40] To be successful, MR requires an IV associated with disease and a causal relationship hypothesis. For diseases in which individually associated variants have small effects, the application of MR may be limited. The ability to exploit genetic variation to mimic randomized controlled trials is a powerful paradigm that can facilitate hypothesis-free causal inference in epidemiology.[38] However, differentiating between a causal relationship for the IV and a pleiotropic relationship can be challenging. Nonetheless, we expect this to be an area of research concentration in the next few years, exploring, eg, the relationship between infection, vitamin D, cannabis use, stress, or metabolic syndrome vs schizophrenia.

Conclusions

The pace of discovery has quickened in schizophrenia genetics. The findings are providing important insights into the nature of schizophrenia risk. For example, with respect to the role of common variants, the polygene profile score for schizophrenia has demonstrated that the risk-associated SNPs that contribute to the score are ubiquitous, and that the total polygene profile score in the general population has a continuous distribution. It is increasingly clear that diverse clinical disorders share genetic risk factors.[41] These developments have important implications in how we will combine genetic and environmental information. With respect to rare structural variants, power to explore the impact of environmental factors in addition to genetic risk will require very large samples. Importantly, the field will need to collate large, population-based samples that not only are genotyped and accurately phenotyped but also have detailed information on candidate exposures. Just as the costs of genotyping are falling, we need to develop cost-effective and high-throughput ways to assess phenotypes and environmental risk factors (ie, the "exposome").[42] Innovative statistical models will be required to handle these data. By combining both genetic and environmental data in research designs, we optimize our chances of mapping the risk landscape of schizophrenia and identifying better treatments and prevention for this poorly understood group of disorders.[2]

Funding

Australian Research Council (FT0991360 to N.R.W.); Australian National Health and Medical Research Council (613672, 613601, 1011506, 1048853 to P.M.V.; 1007677 to J.J.M.; 613602, 613608 to N.R.W.); Queensland Health (to J.J.M.); Stanley Medical Research Institute (to P.B.M.); European Research Council Advanced (to P.B.M.); Lundbeck Foundation Denmark (to P.B.M.).

Acknowledgment

The authors have declared that there are no conflicts of interest in relation to the subject of this study.

References

1. Corvin A. Schizophrenia at a genetics crossroads: where to now? *Schizophr Bull*. 2013;39:490–495.
2. McGrath JJ, Lawlor DA. The search for modifiable risk factors for schizophrenia. *Am J Psychiatry*. 2011;168:1235–1238.
3. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447:661–678.
4. Bergen SE, O'Dushlaine CT, Ripke S, et al. Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol Psychiatry*. 2012;17:880–886.
5. Humphreys K, Grankvist A, Leu M, et al. The genetic structure of the Swedish population. *PLoS One*. 2011;6:e22547.
6. Borglum AD, Demontis D, Grove J, et al. Genome-wide study of association and interaction with maternal cytomegalovirus infection suggests new schizophrenia loci. *Mol Psychiatry*. 2013. doi: 10.1038/mp.2013.2.
7. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era–concepts and misconceptions. *Nat Rev Genet*. 2008;9:255–266.
8. Lichtenstein P, Björk C, Hultman CM, Scolnick E, Sklar P, Sullivan PF. Recurrence risks for schizophrenia in a Swedish national cohort. *Psychol Med*. 2006;36:1417–1425.
9. Yang J, Visscher PM, Wray NR. Sporadic cases are the norm for complex disease. *Eur J Hum Genet*. 2010;18:1039–1043.
10. Do CB, Hinds DA, Francke U, Eriksson N. Comparison of family history and SNPs for predicting risk of complex disease. *PLoS Genet*. 2012;8:e1002973.
11. Mortensen PB, Pedersen MG, Pedersen CB. Psychiatric family history and schizophrenia risk in Denmark: which mental disorders are relevant? *Psychol Med*. 2010;40:201–210.
12. Kendler KS. The impact of varying diagnostic thresholds on affected sib pair linkage analysis. *Genet Epidemiol*. 1988;5:407–419.
13. Agerbo E, Mortensen PB, Wiuf C, et al. Modelling the contribution of family history and variation in single nucleotide polymorphisms to risk of schizophrenia: a Danish national birth cohort-based study. *Schizophr Res*. 2012;134:246–252.
14. van Os J, Pedersen CB, Mortensen PB. Confirmation of synergy between urbanicity and familial liability in the causation of psychosis. *Am J Psychiatry*. 2004;161:2312–2314.
15. Genetic Risk and Outcome in Psychosis (GROUP) Investigators. Evidence that familial liability for psychosis is expressed as differential sensitivity to cannabis: an analysis of patient-sibling and sibling-control pairs. *Arch Gen Psychiatry*. 2011; 68:138–147.
16. van Winkel R; Genetic Risk and Outcome in Psychosis (GROUP) Investigators. Family-based analysis of genetic variation underlying psychosis-inducing effects of cannabis: sibling analysis and proband follow-up. *Arch Gen Psychiatry*. 2011; 68:148–157.
17. Wigman JT, van Winkel R, Ormel J, Verhulst FC, van Os J, Vollebergh WA. Early trauma and familial risk in the development of the extended psychosis phenotype in adolescence. *Acta Psychiatr Scand*. 2012;126:266–273.
18. Svensson AC, Lichtenstein P, Sandin S, Öberg S, Sullivan PF, Hultman CM. Familial aggregation of schizophrenia: the moderating effect of age at onset, parental immigration, paternal age and season of birth. *Scand J Public Health*. 2012;40:43–50.
19. Lenroot RK, Schmitt JE, Ordaz SJ, et al. Differences in genetic and environmental influences on the human cerebral cortex associated with development during childhood and adolescence. *Hum Brain Mapp*. 2009;30:163–174.
20. Bergen SE, Gardner CO, Kendler KS. Age-related changes in heritability of behavioral phenotypes over adolescence and young adulthood: a meta-analysis. *Twin Res Hum Genet*. 2007;10:423–433.

21. Zammit S, Owen MJ, Lewis G. Misconceptions about gene-environment interactions in psychiatry. *Evid Based Ment Health*. 2010;13:65–68.
22. Zammit S, Lewis G, Dalman C, Allebeck P. Examining interactions between risk factors for psychosis. *Br J Psychiatry*. 2010;197:207–211.
23. Duncan LE, Keller MC. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *Am J Psychiatry*. 2011;168:1041–1049.
24. van Winkel R, Henquet C, Rosa A, et al. Evidence that the COMT(Val158Met) polymorphism moderates sensitivity to stress in psychosis: an experience-sampling study. *Am J Med Genet B Neuropsychiatr Genet*. 2008; 147B:10–17.
25. van Winkel R; Genetic Risk and Outcome of Psychosis (GROUP) Investigators. Family-based analysis of genetic variation underlying psychosis-inducing effects of cannabis: sibling analysis and proband follow-up. *Arch Gen Psychiatry*. 2011;68:148–157.
26. van Winkel R, Henquet C, Rosa A, et al. Evidence that the COMT(Val158Met) polymorphism moderates sensitivity to stress in psychosis: an experience-sampling study. *Am J Med Genet B Neuropsychiatr Genet*. 2008;147B:10–17.
27. Lee SH, DeCandia TR, Ripke S, et al. Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ); International Schizophrenia Consortium (ISC); Molecular Genetics of Schizophrenia Collaboration (MGS). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet*. 2012;44:247–250.
28. Wray NR, Visscher PM. Narrowing the boundaries of the genetic architecture of schizophrenia. *Schizophr Bull*. 2010;36:14–23.
29. Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet*. 2012;13:537–551.
30. Sullivan P; 96 Psychiatric Genetics Investigators. Don't give up on GWAS. *Mol Psychiatry*. 2012;17:2–3.
31. Kohane IS, Drazen JM, Campion EW. A glimpse of the next 100 years in medicine. *N Engl J Med*. 2012;367:2538–2539.
32. Perlis RH, Iosifescu DV, Castro VM, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med*. 2012;42:41–50.
33. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res*. 2007;17:1520–1528.
34. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009;460:748–752.
35. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010;42:565–569.
36. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet*. 2011;88:294–305.
37. Wheatley K, Gray R. Commentary: Mendelian randomization–an update on its use to evaluate allogeneic stem cell transplantation in leukaemia. *Int J Epidemiol*. 2004;33:15–17.
38. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. 2003;32:1–22.
39. Timpson NJ, Lawlor DA, Harbord RM, et al. C-reactive protein and its role in metabolic syndrome: mendelian randomisation study. *Lancet*. 2005;366:1954–1959.
40. Bjørngaard JH, Gunnell D, Elvestad MB, et al. The causal role of smoking in anxiety and depression: a Mendelian randomization analysis of the HUNT study. *Psychol Med*. 2013;43:711–719.
41. Smoller JW, Craddock N, Kendler K, et al. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*. 2013;381:1371–1379.
42. Rappaport SM. Discovering environmental causes of disease. *J Epidemiol Community Health*. 2012;66:99–102.