Published in final edited form as:
Am J Psychiatry. 2019 October 01; 176(10): 846–855. doi:10.1176/appi.ajp.2019.18091085.

Penetrance and Pleiotropy of Polygenic Risk Scores for Schizophrenia in 106,160 Patients Across Four Health Care Systems

Amanda B. Zheutlin, Ph.D, **Jessica Dennis, Ph.D**, **Richard Karlsson Linnér, Ph.D**, **Arden Moscati, Ph.D**, **Nicole Restrepo, Ph.D**, **Peter Straub, M.S**, **Douglas Ruderfer, Ph.D**, **Victor M. Castro**, **Chia-Yen Chen, Sc.D**, **Tian Ge, Ph.D**, **Laura M. Huckins, Ph.D**, **Alexander Charney, M.D**, **H. Lester Kirchner, Ph.D**, **Eli A. Stahl, Ph.D**, **Christopher F. Chabris, Ph.D**, **Lea K. Davis, Ph.D**, **Jordan W. Smoller, M.D., Sc.D**

Psychiatric and Neurodevelopmental Genetics Unit (Zheutlin, Chen, Ge, Smoller) and Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston (Chen); Stanley Center for Psychiatric Research, Broad Institute,Cambridge,Mass. (Zheutlin, Chen, Stahl, Smoller); Division of Genetic Medicine, Department of Medicine (Dennis, Straub, Ruderfer, Davis), Vanderbilt Genetics Institute (Dennis, Straub, Ruderfer, Davis), and Department of Biomedical Informatics (Ruderfer), Vanderbilt University Medical Center, Nashville; Department of Economics, School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam (Karlsson Linnér); Autism and Developmental Medicine Institute, Geisinger, Lewisburg, Pa. (Karlsson Linnér, Chabris); Charles Bronfman Institute for Personalized Medicine (Moscati), Pamela Sklar Division of Psychiatric Genomics (Huckins, Charney, Stahl), and Department of Genetics and Genomic Sciences (Huckins, Charney, Stahl,), Icahn School of Medicine at Mount Sinai, New York; Department of Biomedical and Translational Informatics, Geisinger, Rockville, Md. (Restrepo, Kirchner); Research Information Science and Computing, Partners HealthCare, Somerville, Mass. (Castro).

Abstract

Objective: Individuals at high risk for schizophrenia may benefit from early intervention, but few validated risk predictors are available. Genetic profiling is one approach to risk stratification that has been extensively validated in research cohorts. The authors sought to test the utility of this approach in clinical settings and to evaluate the broader health consequences of high genetic risk for schizophrenia.

Methods: The authors used electronic health records for 106,160 patients from four health care systems to evaluate the penetrance and pleiotropy of genetic risk for schizophrenia. Polygenic risk scores (PRSs) for schizophrenia were calculated from summary statistics and tested for association with 1,359 disease categories, including schizophrenia and psychosis, in phenome-wide association studies. Effects were combined through meta-analysis across sites.

Results: PRSs were robustly associated with schizophrenia (odds ratio per standard deviation increase in PRS, 1.55; 95% CI=1.4, 1.7), and patients in the highest risk decile of the PRS

correspondence to Dr. Smoller (jsmoller@mgh.harvard.edu) and Dr. Davis (lea.k.davis@vumc.org).
Drs. Davis and Smoller share last authorship.

distribution had up to 4.6-fold higher odds of schizophrenia compared with those in the bottom decile (95% CI=2.9, 7.3). PRSs were also positively associated with other phenotypes, including anxiety, mood, substance use, neurological, and personality disorders, as well as suicidal behavior, memory loss, and urinary syndromes; they were inversely related to obesity.

Conclusions: The study demonstrates that an available measure of genetic risk for schizophrenia is robustly associated with schizophrenia in health care settings and has pleiotropic effects on related psychiatric disorders as well as other medical syndromes. The results provide an initial indication of the opportunities and limitations that may arise with the future application of PRS testing in health care systems.

Psychiatric disorders are common, and they are responsible for an enormous burden of suffering (1, 2). Globally, approximately 18% of individuals suffer from mental illness every year (3), and 44.7 million of those affected live in the United States (4). Early detection and intervention for serious mental illness is associated with improved outcomes (5–8). However, few reliable predictors of risk or clinical outcomes have been identified. Given the substantial heritability of many psychiatric disorders (9) and their polygenic architecture (10), there is increasing interest in using quantitative measures of genetic risk for risk stratification (11). Polygenic risk scores (PRSs) in particular are easy and inexpensive to generate and can be applied well before illness onset, making them a promising candidate for clinical integration (12). A recent study investigating the clinical utility of PRSs for several common nonpsychiatric diseases (13) found that these scores can identify a larger fraction of high-risk individuals than are identified by clinically validated monogenic mutations; the study authors call explicitly for evaluations of these scores in clinical settings.

To date, PRSs for neuropsychiatric disorders have primarily been evaluated in highly ascertained research samples. Typically, case subjects have obtained a diagnosis through lengthy clinician interviews, and control subjects have no psychiatric history (“clean” case and control samples). In order to bring PRSs to the clinic, however, they must first be demonstrated to have associations with diagnoses in real-world clinical settings, where data are often much messier. Among psychiatric disorders, schizophrenia is perhaps the best candidate for future clinical integration of PRS profiling, as it is highly heritable, has the best-performing PRS among psychiatric disorders in terms of proportion of phenotypic variance explained (7%) (14), and may benefit from early detection and intervention (5–8). Accordingly, we selected the schizophrenia PRS for the present study, as it is the most viable test case for eventual clinical validation of a psychiatric PRS.

We recently established the PsycheMERGE consortium within the National Institutes of Health–funded Electronic Medical Records and Genomics (eMERGE) Network (15, 16) to leverage electronic health record (EHR) data linked to genomic data to facilitate psychiatric genetic research (17). In this first report from PsycheMERGE, we evaluate the performance of a schizophrenia PRS generated from summary statistics published by the Psychiatric Genomics Consortium (PGC) (14) using EHR data on more than 100,000 patients from four large health care systems (Geisinger Health System, Mount Sinai Health System, Partners HealthCare System, and Vanderbilt University Medical Center). We assessed the relative and absolute risk for schizophrenia among individuals at the highest level of genetic risk and considered the clinical utility of the PRS for risk stratification. We also examined pleiotropic effects of the schizophrenia PRS with real-world clinical data by conducting a phenome-wide association study (PheWAS) of 1,359 disease categories.

Finally, we conducted follow-up analyses to characterize the nature of the pleiotropic effects of the schizophrenia PRS. Cross-phenotype associations of polygenic liability to schizophrenia may occur in at least two scenarios (18). In the first, “biological pleiotropy,” the PRS contributes independently to multiple phenotypes. In the second, “mediated pleiotropy,” the PRS increases liability to a second disorder that occurs as a consequence of schizophrenia itself. For example, an association between schizophrenia polygenic risk and diabetes could occur because individuals diagnosed with schizophrenia are more likely both to have elevated schizophrenia PRS and to receive prescriptions for antipsychotic medications, which may result in weight gain and increased liability to diabetes. In this case, the observed relationship between schizophrenia risk and diabetes is mediated by the use of antipsychotic medication. Such scenarios may be difficult to completely disentangle. However, here we use individual-level EHR data to determine whether associations with genetic risk for schizophrenia persist after conditioning on a clinical diagnosis of schizophrenia, related psychosis, or prescription of antipsychotic medication.

METHODS

Hospital-Based Biobanks

Patients who consented to participate in one of four large health care system–based biobanks —the MyCode Community Health Initiative at the Geisinger Health System (GHS) (19), the BioMe Biobank at the Mount Sinai School of Medicine (MSSM) (20), the Partners HealthCare System (PHS) biobank (21), and the Vanderbilt University Medical Center (VUMC) biobank (BioVU) (22)—and had available EHR and genotype data were included in these analyses. At each site, patients were recruited from the general health care system population without systematic recruitment for any particular disease or diagnosis. It is well known that PRSs calculated from genome-wide association studies (GWASs) performed primarily in one ancestry demonstrate poorer performance in other ancestries as a function of differing linkage disequilibrium (LD) structures with causal variants and lack of diversity on genotyping platforms (23, 24). Thus, this study was limited to patients of European-American ancestry for whom genetic data were available that met standard quality control thresholds (see below). Besides these data availability and ancestry filters, no further inclusion or exclusion criteria were applied. Our final sample included 44,436 patients from GHS, 9,569 patients from MSSM, 18,461 patients from PHS, and 33,694 patients from VUMC (a total of 106,160 patients). All patients gave informed consent for biobank research, for which institutional review board approval was obtained at each site.

Quality Control of Genetic Data

Samples were genotyped, imputed, and cleaned at each site individually, as described in detail in the Supplementary Methods section of the online supplement. Quality control procedures at each site followed a similar standard pipeline. DNA from blood samples obtained from biobank participants was assayed using Illumina bead arrays (Omni-Express

Exome, Global Screening, MEGA, MEGA^{EX}, or MEG BeadChips) containing approximately 700,000 to 2 million markers. Samples at each site were genotyped in multiple batches; indicators for genotyping platform and batch were included as covariates in the analyses. As described in the Supplementary Methods section of the online supplement, single-nucleotide polymorphisms (SNPs) were excluded using filters for call rate, minor allele frequency, and heterozygosity at a minimum. Individuals were excluded for excessive missing data or sex errors; a random individual from any pair of related individuals was also excluded (kinship coefficient >0.2). Principal components or self-reported ancestry was used to identify individuals of European ancestry. SNPs that passed the initial phase of quality control were imputed and then converted to best-guess genotypes where only high-quality markers were retained. Ten principal components were generated within the European sample to use as ancestry covariates in all subsequent analyses.

Polygenic Risk Scores

To quantify genetic risk for schizophrenia, we calculated PRSs using summary statistics from the PGC GWAS of schizophrenia (14), which included odds ratios for 9,444,230 variants. PRSs were calculated by two methods, a simple and widely used approach in which SNPs are pruned on the basis of LD and association p values, and a Bayesian approach that can increase accuracy by directly modeling LD structure and adjusting SNP weights accordingly.

LD-pruned PRSs.—We excluded rare variants (minor allele frequency <1%) and variants on the X chromosome, and then, at each site, we clumped SNPs on the basis of association p value (the variant with the smallest p value within a 250-kb range was retained, and all those in LD [r2>0.1] were removed). The resulting SNP lists included 146,464 variants at GHS, 79,837 variants at MSSM, 166,477 variants at PHS, and 229,355 variants at VUMC. Using all available variants (i.e., using a p threshold of 1.0 for inclusion), we generated PRSs for each individual by summing all risk-associated variants weighted by the log odds ratio for that allele from the GWAS. PRSs were converted to z-scores within each health care system to standardize effects across all sites. LD pruning and PRS generation were done using PRSice (25).

Bayesian PRSs.

We used PRS-CS, a Bayesian polygenic prediction method, as an alternative approach for PRS calculation. PRS-CS places a continuous shrinkage (CS) prior on SNP effect sizes and infers posterior SNP weights using GWAS summary statistics and an external LD reference panel (1000 Genomes Project European samples; N=503). PRS-CS allows multivariate modeling of local LD patterns and is robust to diverse underlying genetic architectures, and thus can increase the accuracy of PRS over conventional approaches (26). At each site, weights for all imputed SNPs present on the 1000 Genomes reference panel and the HapMap3 panel were estimated using PRS-CS, resulting in 833,502 available SNPs at GHS, 971,463 at MSSM, 833,502 at PHS, and 604,645 at VUMC. The global shrinkage parameter in the CS prior was fixed at 1 to reflect the highly polygenic genetic architecture of schizophrenia. We generated PRSs for each individual by summing all risk-associated variants weighted by the posterior effect size inferred by PRS-CS for that allele and then converted PRSs to z-scores within each health care system. A Python package for PRS-CS is available on the GitHub repository (https://github.com/getian107/PRScs). PRSs were calculated using PLINK, version 1.9 (27).

EHR-Derived Phenotypes

EHRs contain thousands of diagnostic billing codes from ICD-9 and ICD-10, arranged hierarchically. For example, ICD9:295 is “schizophrenic disorders,” ICD9:295.1 is “disorganized type schizophrenia,” and ICD9:295.12 is “disorganized type schizophrenia, chronic state”; in total, the ICD9: 295 category contains 71 individual ICD-9 codes. To define case status for a variety of diseases, we extracted all ICD-9 and ICD-10 codes available for participating subjects and grouped codes into 1,860 disease categories (called “phecodes”) using a previously developed and validated hierarchical structure (28, 29). For “schizophrenia and other psychotic disorders,” for example, 89 individual ICD-9 codes—all 71 ICD-9 295 codes and 18 related codes (e.g., 298.9, unspecified psychosis)—and 22 ICD-10 codes were mapped to this disease category.

Case and control subjects were designated for each phecode. Individuals with two or more relevant ICD-9/10 codes were considered case subjects, those with zero relevant codes were considered control subjects, and individuals with only one code were excluded (30). To allow analyses of phenome-wide diagnoses that may have varying onset ages, we did not restrict the age range of participants. The proportion of patients (case and control subjects) included in a given PRS-phecode association varied depending on the prevalence of single-code individuals, but the median at each site ranged from 98% to 100%. Phecodes for which there were fewer than 100 cases were excluded from the PheWAS.

Statistical Analysis

Penetrance of schizophrenia PRS in health care systems.—To assess the penetrance of schizophrenia PRS, we measured absolute risk (case prevalence as a function of PRS) and relative risk (odds ratios for the top decile of schizophrenia PRS relative to the remaining population, as well as to the bottom decile) for schizophrenia (phecode 295.1) and psychotic disorders (phecode 295). Odds ratios were calculated at each site for both PRS methods (LD-pruned and Bayesian), regardless of the number of available cases, and then the log odds ratios were combined through fixed-effect inverse variance-weighted meta-analysis using the metafor R package (https://cran.r-project.org/web/packages/metafor/).

Schizophrenia PRS PheWAS.—We conducted PheWASs for both PRS methods in each of the four health care systems, using all phecodes with sufficient sample size (at least 100 cases). Logistic regressions between schizophrenia PRSs and each phecode were run with 10 ancestry principal components, median age within the medical record calculated for each individual using all of their records in the EHR, sex, genotyping platform, and genotyping batch (when available) included as covariates, using the PheWAS R package (29). We used a Bonferroni correction for establishing statistical significance based on the number of phecodes tested at each site. We then meta-analyzed PheWAS effects across health care systems within a given PRS method with a fixed-effect inverse variance-weighted model using the PheWAS R package. Phecodes significantly associated with schizophrenia PRS in the PheWAS meta-analysis were carried forward for a follow-up analysis in which we quantified the risk of the phecode at the extremes of the PRS distribution at each site. Effects were combined across sites through meta-analysis using the metafor R package.

Sensitivity analyses to assess secondary effects of schizophrenia.—To explore whether pleiotropic effects of the schizophrenia PRS were mediated by the diagnosis of schizophrenia itself or by the prescription of antipsychotic medication (the most common treatment for schizophrenia), we conducted four follow-up PheWAS analyses. Given the similarity of primary PheWAS results from the two PRS methods, sensitivity analyses were conducted using the LD-pruned PRS method only. For each follow-up analysis, a PheWAS analysis was conducted as above, with only one of the four following alterations: an additional covariate for diagnosis of psychotic disorders (phecode 295; the broadest schizophrenia-related phecode), an additional covariate for any prescriptions of antipsychotic medication, removing psychosis cases (phecode 295), and removing patients with any antipsychotic medication prescription history.

RESULTS

Our sample included 106,160 patients (56% of them female), across four large U.S. health care systems, who had collectively received over 35 million ICD-9/10 billing codes. The median length of the EHR at each site ranged from 8 to 15 years, and patients' median number of unique visits ranged from 52 to 142 (Table 1).

Penetrance of Schizophrenia PRS in Health Care Systems

PRSs were robustly associated with schizophrenia (phecode 295.1) in the cross-site meta-analysis (odds ratio per standard deviation increase in PRS, 1.55; 95% CI=1.4, 1.7, $p=4.483\times10^{-16}$) (see Table S1 in the online supplement); similar effects were observed using the Bayesian PRS (see Table S2 in the supplement), as well as in each individual health care system (see Tables S3 and S4 in the supplement). Absolute risk for schizophrenia among patients in the top decile was 0.8% (Figure 1), equating to 1.9-fold higher odds of schizophrenia compared with those below the 90th percentile (95% CI=1.5, 2.4, $p=7.81\times10^{-8}$) and 3.3-fold higher odds compared with those in the bottom decile (95% CI=2.1, 5.2, $p=1.16\times10^{-7}$) (Figure 2, Table 2) Similarly, for the Bayesian PRS, absolute risk for the decile was 1.0% (see Figure 1), with an odds ratio of 2.3 compared with the bottom 90th percentile (95% CI=1.9,2.9,$p=1.98\times10^{-14}$) and 4.6 compared with the bottom decile (95% CI=2.9, 7.3, $p=1.37\times10^{-10}$) (Table 2; see also Figure S1 in the online supplement).

Schizophrenia PRS PheWAS

After excluding codes for which no site had at least 100 cases, we conducted a PheWAS using 1,359 disease categories for two PRS methods. The cross-site LD-pruned PRS PheWAS meta-analysis yielded significant associations between schizophrenia PRSs and 29 medical phenotypes, including schizophrenia (Figure 3; see also Table S1). Similar results were observed using the Bayesian PRS (see Table S2) and at each site (see Tables S3 and S4). The strongest cross-site associations were with psychiatric phenotypes for which positive genetic correlations with schizophrenia have been reported, including bipolar disorder, depression, substance use disorders, and anxiety disorders (9). We additionally found associations with personality disorders, suicidal behavior, neurological disorders, memory loss, viral hepatitis, urinary syndromes, and nonspecific somatic symptoms. Obesity and synovitis were inversely associated with schizophrenia PRS. Effect sizes for all significant phenotypes are plotted in Figure 2 and Figure S2 in the online supplement.

Sensitivity Analyses to Assess Secondary Effects of Schizophrenia

We explored whether some of the observed associations might be mediated through a clinical diagnosis of schizophrenia or antipsychotic medication use through a series of sensitivity analyses. Nearly all associations remained significant across all sensitivity analyses (see Table S5 and Figure S2 in the online supplement), although for some phecodes, there was minor variability. Nonetheless, in every analysis, phecodes related to anxiety disorders, mood disorders, substance use disorders, obesity phenotypes, urinary syndromes, and malaise and fatigue remained significant. Associations with suicidal behavior, personality disorders, neurological disorders, memory loss, synovitis and tenosynovitis, and viral hepatitis were less robust, although they remained top phenotypes consistently (see Table S5).

DISCUSSION

We investigated the impact of genetic risk for schizophrenia across the medical phenome in 106,160 patients from four large U.S. health care systems. Several findings from our analyses are noteworthy. First, externally derived PRSs for schizophrenia robustly detected risk for diagnosis of schizophrenia (phecode 295.1) in real-world health care settings (p values,4.48×10^{-16}). The effect sizes (see Table 2) were similar to those observed for corresponding PRSs for atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and many common cancers (13, 31). Second, we leveraged the phenome-wide data available in EHRs to conduct the first psychiatric PRS PheWAS in multiple U.S. health care systems, revealing a range of pleiotropic relationships.

While we observed strong associations with schizophrenia, the effect sizes were more modest than those reported in schizophrenia case-control cohorts ascertained for research purposes. For example, in the original report by the PGC from which the risk scores were derived, depending on the sample, individuals in the top decile of schizophrenia PRS relative to the bottom decile had 7.8-fold to 20.3-fold greater odds of schizophrenia (14), whereas we observed odds ratios of 3.3 and 4.6, depending on the PRS method (see Table 2). There are several potential reasons for this discrepancy. First, case subjects in the PGC meta-analysis met relatively stringent criteria based on clinical interviews by trained research personnel, and control ascertainment often included screening for history of psychiatric or neurological disorders. This approach, typical for research samples, maximizes power for genetic discovery by extreme sampling from the tails of the genetic liability distribution. In contrast, our study used passively collected clinical data—participants were not asked to do anything outside of routine clinical care, thus reducing barriers to participation—and we did not set a clinical symptom threshold for control subjects (other than that they could not be case subjects). This approach more closely approximates an epidemiological design, similar to health registry–based studies in European countries. Thus, although the effect size we observed is likely attenuated as a result of some degree of misclassification, it may better reflect results that would be seen in real-world clinical settings where PRSs were applied to a broad health care population, with little a priori knowledge of clinical symptoms. In addition, we did not restrict the age range of case and control subjects, which may have further reduced the apparent effect size of the schizophrenia PRS (some individuals in our sample who have not yet reached the age of illness onset may have been misclassified as control subjects).

Although the PRS effects we observed were not large enough on their own to stratify risk in a clinical setting (i.e., to discriminate between cases and controls on an individual level with high accuracy), they are comparable to those of risk factors in established risk calculators. For example, two well-established coronary artery disease risk factors, smoking and diabetes, were estimated in the Framingham Heart Study to have hazard ratios #2.0 (32)—similar to the observed risk for the top schizophrenia PRS decile here. In light of this, we speculate that incorporating genetic risk could have an impact in psychiatry, especially as enhanced performance may be possible through a variety of means. For example, we implemented two PRS methods, a standard LD-pruning approach and a newer Bayesian one, to evaluate the robustness and consistency of our results. While the differences in results were not large, the Bayesian method produced larger effect estimates overall, including for schizophrenia (see Table 2). These findings support the use of newer risk scoring methods that can incorporate more genetic variants by directly modeling LD structure. The precision of PRSs may also increase through larger discovery sample sizes (12) and with refinement of EHR-based case definitions.

Nonetheless, it remains to be seen whether combining PRS risk estimates with other clinical predictors can meaningfully contribute to individualized risk assessment in psychiatry. As expected, the areas under the receiver operating curve—a common metric used to evaluate predictive performance—for PRS alone were modest (0.60–0.71 across sites; see Table S6 in the online supplement), although they were similar to those observed for schizophrenia PRSs in research samples (0.59–0.81) (14), as well as similarly computed PRSs for other complex diseases, including type 2 diabetes (0.70), breast cancer (0.66), and inflammatory bowel disease (0.60) (13). A remaining challenge for all risk stratification efforts in low-prevalence diseases (such as schizophrenia) is that even at high risk thresholds (e.g., the top 10% of a PRS), most individuals are not affected, limiting the utility of stratification for clinical practice. It may be that adequate precision will only be achieved through incorporation of many different measures of risk (e.g., genetic and nongenetic factors).

Schizophrenia PRSs were also associated with broader effects on mental health, including higher risks for anxiety, mood, substance use, personality, and neurological disorders, as well as memory loss and suicidal behavior. Anxiety, mood, and substance use disorders have all previously been linked to genetic risk for schizophrenia (9, 33–35), and our results confirm in a clinical setting that these disorders share genetic risk. Certain personality disorders have also been linked to genetic liability for schizophrenia (36, 37) (e.g., schizotypal or schizoid), and there is some evidence that personality dimensions in adolescence predict future psychopathology, including schizophrenia (38). Similarly, family history of schizophrenia has been associated with suicidal behavior (39). However, results from our sensitivity analyses suggest that the relationships between schizophrenia and neurological disorders, personality disorders, suicidal behavior, and memory loss may be consequences of a schizophrenia diagnosis rather than due to shared genetic risk (see Figure S2 in the online supplement).

Genetic liability for schizophrenia was associated with many nonpsychiatric syndromes as well, including obesity, urinary syndromes, viral hepatitis, synovitis and tenosynovitis, and malaise and fatigue. Intriguingly, obesity and morbid obesity were significantly negatively associated with schizophrenia PRSs (see Table S1 in the online supplement). This is somewhat surprising given the known phenotypic correlation between schizophrenia and obesity (40). Nonetheless, three previous reports found significant inverse genetic correlations between body mass index and schizophrenia (41–43), while a fourth reported an inverse but non-significant relationship (44). This may suggest that elevated rates of obesity among patients with schizophrenia may be a consequence of the disease, potentially due to antipsychotic use or poor support for proper nutrition. We also found an inverse association between genetic liability for schizophrenia and diabetes, but only in sensitivity analyses controlling for a schizophrenia diagnosis or antipsychotic medication history. It may be that this negative genetic correlation was attenuated in the primary analysis (i.e., including patients with schizophrenia and antipsychotic medication history with no statistical control) because of diabetes-promoting effects of antipsychotic medications within the same individuals who were at high genetic risk for schizophrenia (40). In general, pleiotropic effects may have implications for risk communication if PRS testing is deployed in clinical settings in the future.

Our results should be interpreted in light of several limitations. First, because of small numbers of patients of other ancestries, our analyses were restricted to patients of European descent, and the generalizability to individuals of non-European ancestry remains to be determined. Second, our phenotype definitions relied on very simple rules and disregarded many variables of potential importance, including medical history of related disorders, setting of diagnosis (i.e., inpatient or outpatient; physician specialty), and treatment for the disease of interest. This was by design in order to mimic a real-world clinical population in which PRSs may be implemented for clinical decision support; however, the approach is sensitive to misclassifications that occur in a clinical setting. Future work refining case and control definitions using natural language processing algorithms may improve the predictive performance of PRSs and other risk factors for clinically derived phenotypes (45, 46). Third, our results varied to some degree between sites (see Tables S3 and S4 in the online supplement), perhaps most notably for schizophrenia, suggesting that demographic and disease distributions in any given health care system will influence penetrance and pleiotropy. However, we tested for between-site heterogeneity for schizophrenia (phecode 295.1), and although this test has relatively low power, it showed no evidence of significant heterogeneity (p values >0.45). Relatedly, disease prevalence was often lower in the overall health care system relative to the participants enrolled in the biobanks (a subset of those patients) (see Table S7 in the online supplement). In general, case prevalence in the biobanks was more representative of population-level prevalence than it was in the health care systems, suggesting that the discrepancies may be due to biobank patients generally having a longer duration of EHR follow-up and therefore more opportunity to receive a diagnosis than patients in the overall health care system (see Table S7). Finally, although our analyses comprise the largest test of a schizophrenia PRS in EHR data to date, additional phenotypes may show significant association in future larger-scale PheWASs.

CONCLUSIONS

We have demonstrated that an available measure of polygenic risk for schizophrenia is robustly associated with schizophrenia across four large health care systems using EHR data. While the observed penetrance of schizophrenia PRS is attenuated in these settings compared with previous estimates derived from research cohorts, effect sizes are comparable to those seen for risk factors commonly used in clinical settings. We also found that polygenic risk for schizophrenia has pleiotropic effects on related psychiatric disorders as well as several nonpsychiatric symptoms and syndromes. Our results provide an initial indication of the opportunities and limitations that may arise with the future application of PRS testing in health care systems.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported in part by NIMH grant R01MH118233 (to Drs. Smoller and Davis). Dr. Ge was supported in part by National Institute on Aging grant K99AG054573, Dr. Dennis was supported by the Canadian Institutes of Health Research (MFE-142936), Dr. Stahl was supported in part by NIMH grants U01MH109536, R01MH106531, and R01MH095034, Dr. Davis was supported by NIMH grant R01MH113362, Dr. Ruderfer was supported by NIMH grant R01MH111776, and Dr. Smoller was supported in part by National Human Genome Research Institute grant U01HG008685 supporting the eMERGE Network; he is also a Tepper Family MGH Research Scholar and is supported in part by the Demarest Lloyd, Jr. Foundation.

Presented at the World Congress of Psychiatric Genetics, Glasgow, Scotland, October 11–15, 2018.

The VUMC components of the project were conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville. The data sets used for the project were obtained from Vanderbilt University Medical Center's Synthetic Derivative and BioVU, which are supported by numerous sources: institutional funding, private agencies, and federal grants. These include the NIH-funded Shared Instrumentation Grant S10RR025141 and CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975 from the National Center for Advancing Translational Sciences. Additional funding was provided by NIH through grants P50GM115305 and U19HL065962. The Mount Sinai BioMe Biobank is supported by the Andrea and Charles Bronfman Philanthropies.

The authors acknowledge the expert technical support of the VANTAGE and VANGARD core facilities, supported in part by the Vanderbilt-Ingram Cancer Center (P30 CA068485) and Vanderbilt Vision Center (P30 EY08126). The authors also acknowledge the Partners Biobank for providing samples, genomic data, and health information data. The authors thank the MyCode Community Health Initiative participants for their permission to utilize their health and genomics information in the DiscovEHR collaboration. The DiscovEHR study was funded in part by the Regeneron Genetics Center, and the Geisinger Clinic was funded by grant U01HG008679.

Dr. Smoller is an unpaid member of the Bipolar/Depression Research Community Advisory Panel of 23andMe. Dr. Kirchner has received funding from Regeneron Genetics Center as part of the DiscovEHR study. All other authors report no financial relationships with commercial interests.

REFERENCES

1. Walker ER, McGee RE, Druss BG: Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. JAMA Psychiatry 2015; 72:334–341 [PubMed: 25671328]
2. Vos T, Abajobir AA, Abbafati C, et al.: Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 2017; 390:1211–1259 [PubMed: 28919117]
3. Steel Z, Marnane C, Iranpour C, et al.: The global prevalence of common mental disorders: a systematic review and meta-analysis, 1980–2013. Int J Epidemiol 2014; 43:476–493
4. Substance Abuse and Mental Health Services Administration (SAMHSA): Key Substance Use and Mental Health Indicators in the United States: Results From the 2016 National Survey on Drug Use and Health. Rockville, Md, SAMHSA, 2017
5. Albert N, Melau M, Jensen H, et al.: The effect of duration of untreated psychosis and treatment delay on the outcomes of prolonged early intervention in psychotic disorders. NPJ Schizophr 2017; 3:34 [PubMed: 28951544]
6. Tang JYM, Chang WC, Hui CLM, et al.: Prospective relationship between duration of untreated psychosis and 13-year clinical outcome: a first-episode psychosis study. Schizophr Res 2014; 153:1–8 [PubMed: 24529612]
7. Amminger GP, Edwards J, Brewer WJ, et al.: Duration of untreated psychosis and cognitive deterioration in first-episode schizophrenia. Schizophr Res 2002; 54:223–230 [PubMed: 11950547]
8. Wang PS, Berglund P, Olfson M, et al.: Failure and delay in initial treatment contact after first onset of mental disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005; 62: 603–613 [PubMed: 15939838]
9. Brainstorm Consortium, Anttila V, Bulik-Sullivan B, et al.: Analysis of shared heritability in common disorders of the brain. Science 2018; 360:eaap8757
10. Smoller JW, Andreassen OA, Edenberg HJ, et al.: Psychiatric genetics and the structure of psychopathology. Mol Psychiatry 2019; 24:409–420 [PubMed: 29317742]
11. Vassos E, Di Forti M, Coleman J, et al.: An examination of polygenic score risk prediction in individuals with first-episode psychosis. Biol Psychiatry 2017; 81:470–477 [PubMed: 27765268]
12. Zheutlin AB, Ross DA: Polygenic risk scores: what are they good for? Biol Psychiatry 2018; 83:e51–e53 [PubMed: 29759133]
13. Khera AV, Chaffin M, Aragam KG, et al.: Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 2018; 50:1219–1224 [PubMed: 30104762]
14. Schizophrenia Working Group of the Psychiatric Genomics Consortium: Biological insights from 108 schizophrenia-associated genetic loci. Nature 2014; 511:421–427 [PubMed: 25056061]
15. Crawford DC, Crosslin DR, Tromp G, et al.: eMERGEing progress in genomics: the first seven years. Front Genet 2014; 5:184 [PubMed: 24987407]
16. Gottesman O, Kuivaniemi H, Tromp G, et al.: The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013; 15:761–771 [PubMed: 23743551]
17. Smoller JW: The use of electronic health records for psychiatric phenotyping and genomics. Am J Med Genet B Neuropsychiatr Genet 2018; 177:601–612 [PubMed: 28557243]
18. Solovieff N, Cotsapas C, Lee PH, et al.: Pleiotropy in complex traits:challenges and strategies. Nat Rev Genet 2013; 14:483–495 [PubMed: 23752797]
19. Carey DJ, Fetterolf SN, Davis FD, et al.: The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet Med 2016; 18: 906–913 [PubMed: 26866580]
20. Belbin GM, Odgis J, Sorokin EP, et al.: Genetic identification of a common collagen disease in Puerto Ricans via identity-by-descent mapping in a health system. eLife 2017; 6:e25060

21. Karlson EW, Boutin NT, Hoffnagle AG, et al.: Building the Partners Healthcare Biobank at Partners Personalized Medicine: informed consent, return of research results, recruitment lessons, and operational considerations. J Pers Med 2016; 6:E2 [PubMed: 26784234]

22. Danciu I, Cowan JD, Basford M, et al.: Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform 2014; 52:28–35 [PubMed: 24534443]

23. Duncan L, Shen H, Gelaye B, et al.: Analysis of polygenic score usage and performance in diverse human populations. bioRxiv, 11 3, 2018 (doi: 10.1101/398396)

24. Martin AR, Kanai M, Kamatani Y, et al.: Current clinical use of polygenic scores will risk exacerbating health disparities. bioRxiv, 2 1, 2019 (doi: 10.1101/441261)

25. Euesden J, Lewis CM, O'Reilly PF: PRSice: polygenic risk score software. Bioinformatics 2015; 31:1466–1468 [PubMed: 25550326]

26. Ge T, Chen C- Y, Ni Y, et al.: Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun 2019; 10: 1776 [PubMed: 30992449]

27. Chang CC, Chow CC, Tellier LCAM, et al.: Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015; 4:7 [PubMed: 25722852]

28. Wei WQ, Bastarache LA, Carroll RJ, et al.: Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS One 2017; 12:e0175508

29. Carroll RJ, Bastarache L, Denny JCR: R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics 2014; 30:2375–2376 [PubMed: 24733291]

30. Wei WQ, Teixeira PL, Mo H, et al.: Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016; 23(e1):e20–e27 [PubMed: 26338219]

31. Fritsche LG, Gruber SB, Wu Z, et al.: Association of polygenic risk scores for multiple cancers in a phenome-wide study: results from the Michigan Genomics Initiative. Am J Hum Genet 2018; 102: 1048–1061 [PubMed: 29779563]

32. D'Agostino RB Sr, Vasan RS, Pencina MJ, et al.: General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 2008; 117:743–753 [PubMed: 18212285]

33. Gandal MJ, Haney JR, Parikshak NN, et al.: Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. Science 2018; 359:693–697 [PubMed: 29439242]

34. Smoller JW, Craddock N, Kendler K, et al.: Identification of risk loci with shared effects on five major psychiatric disorders: a genomewide analysis. Lancet 2013; 381:1371–1379 [PubMed: 23453885]

35. Mistry S, Harrison JR, Smith DJ, et al.: The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: systematic review. Schizophr Res 2019; 197:2–8

36. Nelson MT, Seal ML, Pantelis C, et al.: Evidence of a dimensional relationship between schizotypy and schizophrenia: a systematic review. Neurosci Biobehav Rev 2013; 37:317–327 [PubMed: 23313650]

37. Bigdeli TB, Bacanu SA, Webb BT, et al.: Molecular validation of the schizophrenia spectrum. Schizophr Bull 2014; 40:60–65 [PubMed: 23970557]

38. Newton-Howes G, Horwood J, Mulder R: Personality characteristics in childhood and outcomes in adulthood: findings from a 30 year longitudinal study. Aust N Z J Psychiatry 2015; 49: 377–386 [PubMed: 25688124]

39. Laursen TM, Trabjerg BB, Mors O, et al.: Association of the polygenic risk score for schizophrenia with mortality and suicidal behavior: a Danish population-based study. Schizophr Res 2017; 184:122–127 [PubMed: 27939829]

40. Annamalai A, Kosir U, Tek C: Prevalence of obesity and diabetes in patients with schizophrenia. World J Diabetes 2017; 8:390–396 [PubMed: 28861176]

41. So HC, Chau KL, Ao FK, et al.: Exploring shared genetic bases and causal relationships of schizophrenia and bipolar disorder with 28 cardiovascular and metabolic traits. Psychol Med 2019; 49:1286–1298 [PubMed: 30045777]

42. Akiyama M, Okada Y, Kanai M, et al.: Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat Genet 2017; 49:1458–1467 [PubMed: 28892062]

43. Duncan LE, Shen H, Ballon JS, et al.: Genetic correlation profile of schizophrenia mirrors epidemiological results and suggests link between polygenic and rare variant (22q11.2) cases of schizophrenia. Schizophr Bull 2018; 44:1350–1361 [PubMed: 29294133]

44. Bulik-Sullivan B, Finucane HK, Anttila V, et al.: An atlas of genetic correlations across human diseases and traits. Nat Genet 2015; 47: 1236–1241 [PubMed: 26414676]

45. 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 [PubMed: 21682950]

46. Castro VM, Minnier J, Murphy SN, et al.: Validation of electronic health record phenotyping of bipolar disorder cases and controls. Am J Psychiatry 2015; 172:363–372 [PubMed: 25827034]

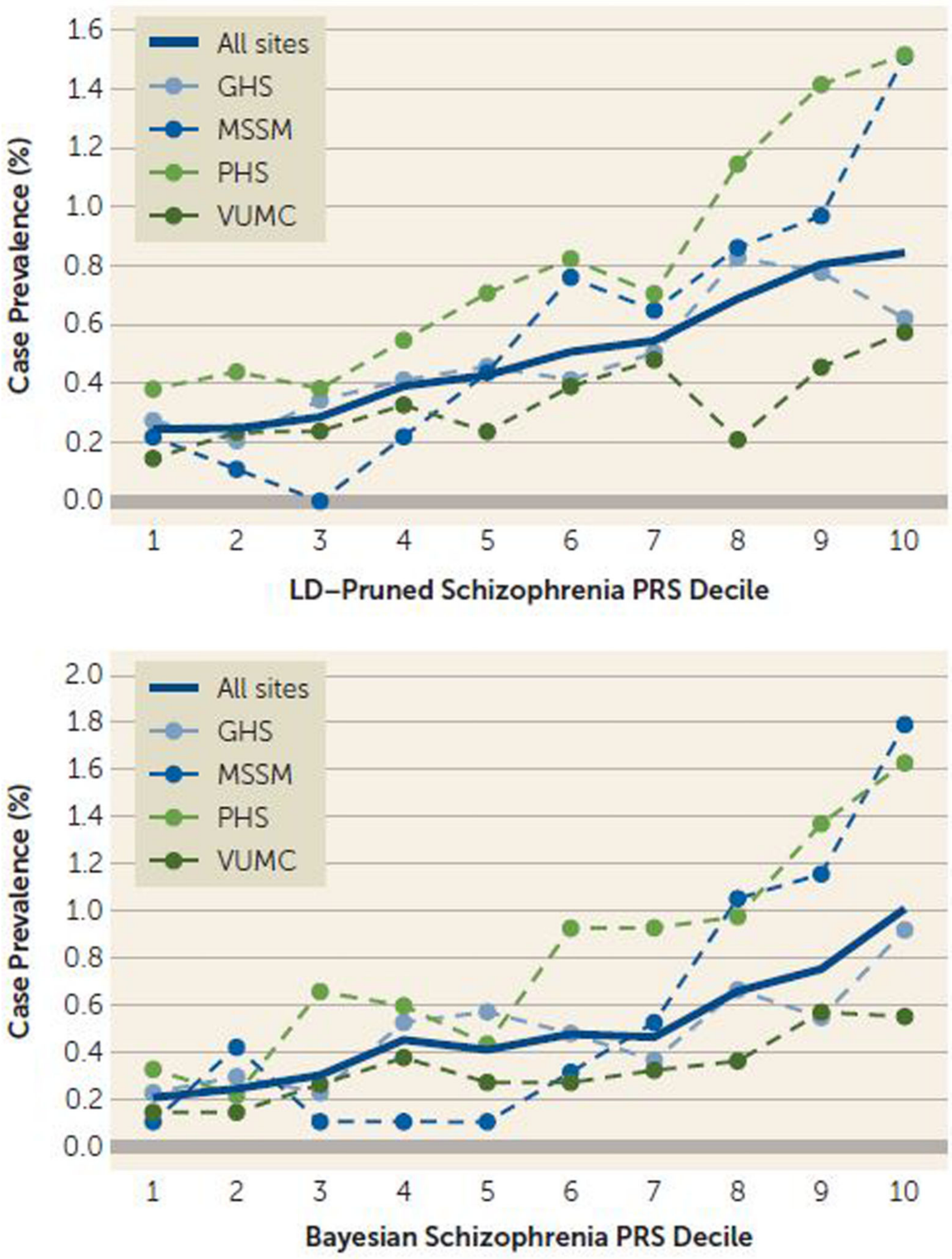

FIGURE 1. Schizophrenia case prevalence by polygenic risk score (PRS) decile among patients in four health care systems[a]

[a] Schizophrenia (phecode 295.1) case prevalence by site and across all health care systems was plotted by schizophrenia PRS decile for both PRS methods. GHS=Geisinger Health System; LD=linkage disequilibrium; MSSM=Mount Sinai School of Medicine; PHS=Partners HealthCare System; VUMC=Vanderbilt University Medical Center.

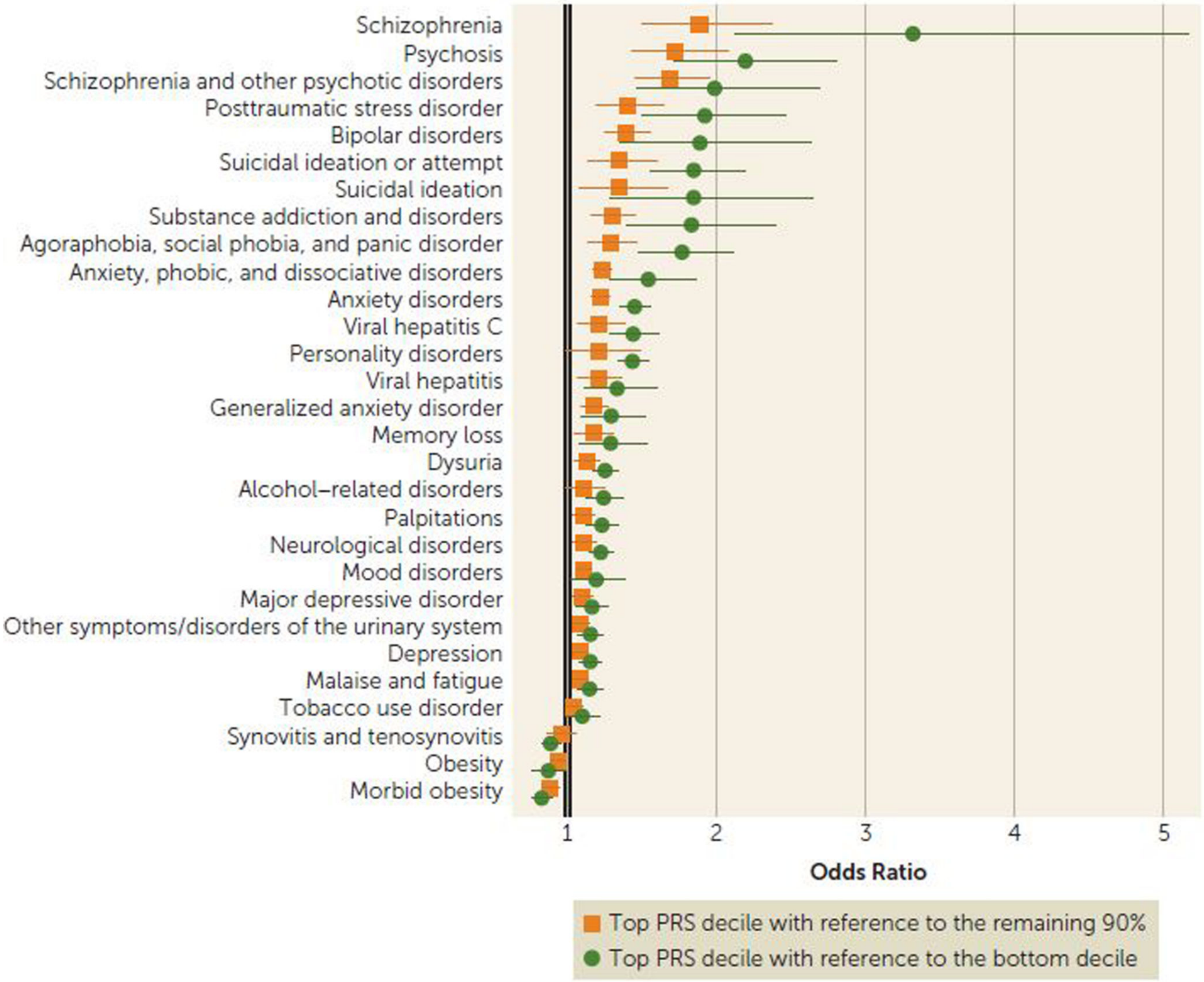

FIGURE 2. Odds ratios for top schizophrenia polygenic risk score (PRS) decile among patients in four health care systems[a]

[a] Odds ratios for phenotypes significant in LD-pruned PRS phenome-wide association study meta-analysis were plotted for the top PRS decile with reference to both the remaining 90% and the bottom decile. The vertical line (odds ratio=1) reflects no change in risk. Error bars indicate 95% confidence intervals. LD=linkage disequilibrium.

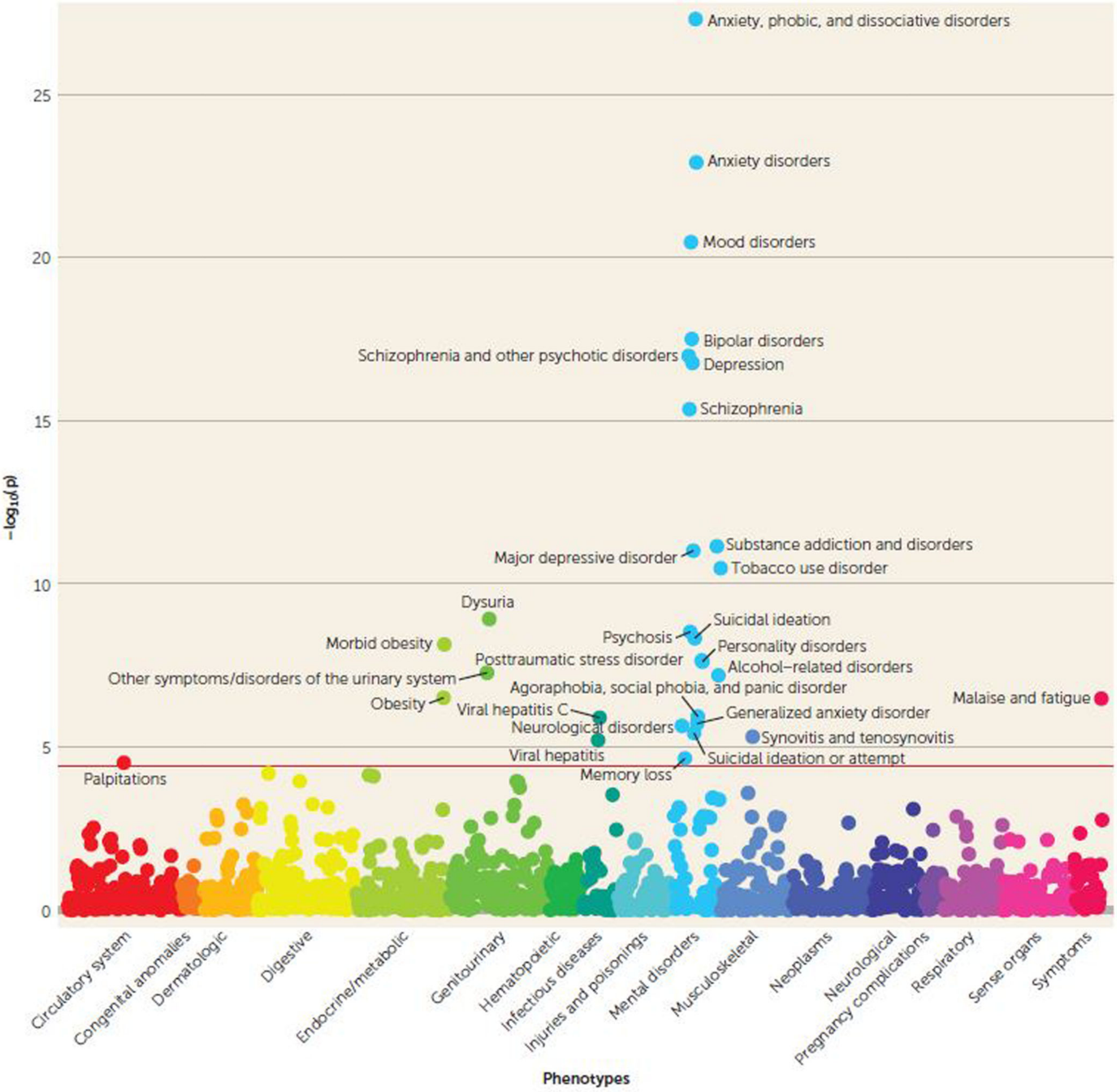

FIGURE 3. Schizophrenia polygenic risk score (PRS) phenome-wide association study meta-analysis across four health care systems[a]

[a] The figure is a Manhattan plot for phenome-wide association with LD-pruned schizophrenia PRSs meta-analyzed across four health care systems (1,359 phenotypes; 106,160 patients). The horizontal axis indicates phenotype (grouped by broad disease category) and the vertical axis indicates the significance (–log10 p; two-tailed) of the association derived by logistic regression. The horizontal red line within the graph indicates phenome-wide-level significance ($p=3.7\times10^{-5}$) using Bonferroni correction, and all phenotypes that pass this threshold are labeled. All significant effects were positive (i.e., higher polygenic risk scores resulted in higher incidence of the phenotype), with three

exceptions: morbid obesity, obesity, and synovitis and tenosynovitis. LD=linkage disequilibrium.

TABLE 1.

Demographic and clinical characteristics of patients in four health care systems in a study of schizophrenia polygenic risk score (N=106,160)[a]

	Site							
Measure	**GHS (N=44,436)**		**MSSM (N=9,569)**		**PHS (N=18,461)**		**VUMC (N=33,694)**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age (years)	60.2	16.9	57.2	19.8	58.5	16.4	57.9	20.0
	N	%	N	%	N	%	N	%
Female	26,094	59	4,955	52	9,913	54	18,089	54
Schizophrenia patients	211	<1	53	<1	148	<1	110	<1
Psychosis patients	499	1	66	<1	385	2	451	1
All ICD-9/10 codes	20,083,326		1,021,072		5,972,131		8,043,419	
Unique ICD-9/10 codes	29,766		16,535		26,441		26,283	
	Median		Median		Median		Median	
ICD-9/10 codes per patient	317		44		184		150	
Visits per patient	142		81		70		52	
Electronic health record length (days)	5,509		2,942		4,729		3,884	

[a] Schizophrenia case subjects had two or more ICD codes included in phecode 295.1; for psychosis case subjects, phecode 295. Age is defined as the patients' age at their most recent hospital visit in which they received an ICD-9/10 code. A visit is both patient- and date-specific but may include many individual ICD-9/10 codes. GHS=Geisinger Health System; MSSM=Mount Sinai School of Medicine; PHS=Partners HealthCare System; VUMC=Vanderbilt University Medical Center.

TABLE 2.

Odds ratios for schizophrenia and psychotic disorders among patients in four health care systems[a]

Disorder Group and PRS Method	Risk Group	Risk Case Prevalence (%)	Reference Group	Reference Case Prevalence (%)	Odds Ratio	95% CI	p
Schizophrenia							
LD-pruned	Top 10%	0.8	Remaining 90%	0.5	1.9	1.5–2.4	7.81310^{-8}
LD-pruned	Top 10%	0.8	Bottom 10%	0.2	3.3	2.1–5.2	1.16310^{-7}
Bayesian	Top 10%	1.0	Remaining 90%	0.4	2.3	1.9–2.9	1.98310^{-14}
Bayesian	Top 10%	1.0	Bottom 10%	0.2	4.6	2.9–7.3	1.37310^{-10}
Psychosis							
LD-pruned	Top 10%	2.1	Remaining 90%	1.3	1.7	1.5–2.0	2.00310^{-12}
LD-pruned	Top 10%	2.1	Bottom 10%	0.9	2.2	1.7–2.8	4.14310^{-10}
Bayesian	Top 10%	2.1	Remaining 90%	1.3	1.6	1.4–1.9	1.75310^{-10}
Bayesian	Top 10%	2.1	Bottom 10%	1.0	2.1	1.6–2.7	2.75310^{-9}

[a] Overall sample case prevalence was 0.5% for schizophrenia (phecode 295.1) and 1.4% for psychosis (phecode 295). LD=linkage disequilibrium; PRS=polygenic risk score.