Penetrance and pleiotropy of polygenic risk scores for schizophrenia in 90,000 patients across three healthcare systems

Amanda B Zheutlin[1], Jessica Dennis[2,3], Nicole Restrepo[4], Peter Straub[2,3], Douglas Ruderfer[2,3,5], Victor M Castro[6], Chia-Yen Chen[1,7,8], H. Lester Kirchner[4], Christopher F. Chabris[9], Lea K Davis[2,3,*], & Jordan W Smoller[1,7,*]

1 – Psychiatric & Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA
2 – Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN
3 – Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN
4 – Department of Biomedical & Translational Informatics, Geisinger, Rockville, MD
5 – Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN
6 – Research Information Science and Computing, Partners HealthCare, Somerville, MA
7 – Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA
8 – Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA
9 – Autism and Developmental Medicine Institute, Geisinger, Lewisburg, PA
* – Co-last authorship

Corresponding authors:
Jordan W Smoller
Psychiatric and Neurodevelopmental Genetics Unit
Massachusetts General Hospital
185 Cambridge St.,
Boston, MA 02114
617-724-0835
jsmoller@mgh.harvard.edu

Lea K Davis
Division of Genetic Medicine, Department of Medicine
Vanderbilt University Medical Center
2215 Garland Ave
Nashville, TN 37232
615-875-9478
lea.k.davis@vumc.org

Abstract

BACKGROUND: Individuals at high risk 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, but its utility in clinical settings remains largely unexplored. Moreover, the broad health consequences of a high genetic risk of schizophrenia are poorly understood, despite being highly relevant to treatment decisions.

METHODS: We used electronic health records of 91,980 patients from three large healthcare systems to evaluate the penetrance and pleiotropy of genetic risk for schizophrenia. Polygenic risk scores (PRSs) for schizophrenia were calculated from meta-analysis summary statistics and tested for association with schizophrenia diagnostic codes and 1338 code-defined disease categories in a phenome-wide association study. Effect estimates were meta-analyzed across sites, and follow-up analyses evaluated the effect of a schizophrenia diagnosis.

RESULTS: PRSs were robustly associated with schizophrenia (odds ratio per standard deviation increase in PRS = 1.65 [95% confidence interval (CI), 1.5-1.8], $p = 1.25 \times 10^{-16}$) and patients in the highest risk decile of the PRS distribution had a four-fold increased odds of schizophrenia compared to those in the bottom decile (95% CI, 2.4-6.5, $p = 4.43 \times 10^{-8}$). PRSs were also associated with other psychiatric phenotypes, including anxiety disorders, bipolar disorder, depression, substance use disorders, personality disorders, and suicidal behavior. Non-psychiatric associations included heart palpitations, urinary syndromes, obesity, and nonspecific somatic symptoms. Most associations remained significant when conditioning on a diagnosis of schizophrenia, indicating genetic pleiotropy.

CONCLUSIONS: We demonstrate that an available measure of genetic risk for schizophrenia is robustly associated with schizophrenia in healthcare settings and has pleiotropic effects on related psychiatric disorders as well as other medical 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 healthcare systems.

Psychiatric disorders are common and responsible for an enormous burden of suffering[1,2]. Approximately 18% of individuals globally suffer from mental illness every year[3], 44.7 million of whom live in the US[4]. Early detection and intervention for serious mental illness is associated with improved outcomes[5–10]. However, few reliable predictors of risk or clinical outcomes have been identified, limiting the ability to offer early intervention or targeted treatment strategies. Given the substantial heritability of many psychiatric disorders[11] and their polygenic architecture[12], there is increasing interest in using quantitative measures of genetic risk for risk stratification[13]. 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[14]. In fact, a recent study investigating the clinical utility of PRSs for several common, non-psychiatric diseases found that PRS can identify a larger fraction of high risk individuals than are identified by clinically-validated monogenic mutations, and called explicitly for evaluations of these scores in clinical settings[15].

To date, PRSs for neuropsychiatric disorders have primarily been validated in highly ascertained research samples. Typically, cases have obtained a diagnosis through expert clinician interviews standardized across the study, and controls have no psychiatric history (i.e., "clean" cases and controls). In order to bring PRSs to the clinic, however, they must first be validated as predictors of clinical diagnoses in real-world clinical settings, where data are often much messier. Among psychiatric disorders, schizophrenia is perhaps the best candidate for clinical integration of PRS profiling as it is highly heritable and a PRS for schizophrenia has been shown to explain a greater proportion of phenotypic variance (7%)[16] compared to those for other psychiatric disorders. Accordingly, we selected the schizophrenia PRS for the present study as it is currently the most viable test case for clinical validation of a psychiatric PRS.

We recently established the PsycheMERGE consortium within the NIH-funded Electronic Medical Records and Genomics (eMERGE) Network[17,18] to leverage electronic health record (EHR) data linked to genomic data to facilitate psychiatric genetic research[19]. In this first report from PsycheMERGE, we evaluate the performance of a schizophrenia PRS generated from summary statistics published by the Psychiatric Genomics Consortium[16] using EHR data on more than 90,000 patients from three large healthcare systems (Partners Healthcare System, Vanderbilt University Medical Center, and Geisinger Health System). 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 1338 disease categories. To our knowledge this is the first effort to combine PheWAS effects across multiple hospital-based biobanks.

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[20]. In the first ("biological pleiotropy"), the PRS contributes independently to multiple phenotypes. In the second scenario ("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 to have both elevated schizophrenia PRS *and* to be prescribed 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 to control clinical symptoms. These scenarios may be difficult to completely disentangle. However, it is possible to use EHR data to determine whether associations with genetic risk for schizophrenia persist after conditioning on a clinical diagnosis of schizophrenia or related psychosis.

Methods

Hospital-based Biobanks

Patients that consented to participate in one of three large healthcare system-based biobanks – the Vanderbilt University Medical Center (VUMC) biobank (BioVU)[21], the MyCode Community Health Initiative at the Geisinger Health System (GHS)[22], or the Partners Healthcare System (PHS) biobank[23] – and had available EHR and genotype data were included in these analyses. Only patients of European-American ancestry with genetic data that met standard quality control thresholds were retained to reduce any effects of population stratification or genotyping error on the subsequent analyses (see Quality Control of Genetic Data). Our final sample included 18,370 patients from VUMC, 56,926 patients from GHS, and 16,684 patients from PHS (91,980 total participants). All patients gave informed consent for biobank research and IRB approval was obtained at each site.

Quality Control of Genetic Data

Samples were genotyped, imputed, and cleaned at each site individually, the details of which are described in Supplementary Methods. However, quality control procedures at each site followed a similar standard pipeline. DNA from blood samples obtained from biobank participants were assayed using Illumina bead arrays (OmniExpress Exome, MEGA, MEGA^{EX}, or MEG BeadChips) containing approximately one to two million markers. Samples at each site were genotyped in multiple batches, and in some cases, batches used different arrays. Indicators for genotyping platform and batch were included as covariates in the analyses. As described in Supplementary Methods, single nucleotide polymorphisms (SNPs) were excluded using filters for call rate, minor allele frequency, violations of Hardy-Weinberg equilibrium, batch effects, and heterozygosity. Individuals were excluded for excessive missing data or sex errors; a random individual from any pair of related individuals was also excluded. Principal components were used to identify individuals of European ancestry. SNPs that passed this initial phase of quality control were imputed using a 1000 Genomes reference panel and then converted to best guess genotypes where only high-quality markers were retained (INFO > .9). Ten principal components were generated within the European sample to use as ancestry covariates in all subsequent analyses.

Polygenic Risk Scores

In order to quantify genetic risk for schizophrenia, we calculated PRSs using summary statistics from the most recent available genome-wide association study (GWAS) of schizophrenia from the Psychiatric Genomics Consortium[16]. These summary statistics included odds ratios (ORs) for 9,444,230 variants; we excluded variants from this list on the X chromosome and, at each site, clumped SNPs based on association p-value (the variant with the smallest p-value within a 250kb range was retained and all those in linkage disequilibrium, $r^2 > .1$, were removed). The resulting SNP lists included 117,774 variants at VUMC, 166,477 at PHS, and 247,698 at GHS. Using all available variants (i.e., using a p-value threshold of 1.0 for inclusion), we generated PRSs for each individual by summing all risk-associated common variants (minor allele frequency >1%) across the genome, each weighted by the log(OR) for that allele from the GWAS. PRSs were converted to z-scores within each healthcare system to standardize effects across all sites. All analyses were run in R, version 3.4.3; LD clumping and PRS generation were done using PRSice[24].

EHR-derived Phenotypes

EHRs contain thousands of diagnostic billing codes from the International Classification of Diseases, 9th and 10th editions (ICD-9/10) which are arranged hierarchically. For example, 295 is 'schizophrenic disorders', 295.1 is 'disorganized type schizophrenia', and 295.12 is 'disorganized type schizophrenia, chronic state'; in total, the 295 category contains 71 individual ICD-9 codes. To define case status for a variety of diseases, we extracted all ICD-9 codes available for participating subjects and grouped codes into 1645 disease categories (called 'phecodes') using a hierarchical structure previously developed and validated[25,26]. For "schizophrenia and other psychotic disorders", for example, 89 individual ICD-9 codes were mapped to this disease category – all 71 ICD-9 295 codes and 18 related codes (e.g., 298.9, unspecified psychosis). Because a validated structure that incorporates ICD-10 codes has not yet been developed, we restricted our analyses to ICD-9 codes.

Cases and controls were designated for each phecode. Individuals with two or more relevant ICD-9 codes were considered a case, those with zero relevant codes were considered a control, and individuals with only one code were excluded[27]. To enable analyses of phenome-wide diagnoses that may have varying ages of onset, we did not restrict the age range of participants. The number of patients (cases and controls) included in the PheWAS varied depending on the prevalence of single-code individuals, but ranged from 79.0%– 99.9% of total patients at each site. Phecodes for which there were fewer than 100 cases were excluded from the PheWAS.

Statistical Analyses

Penetrance of schizophrenia PRS in healthcare systems. To assess the penetrance of schizophrenia PRS, we calculated case prevalence for schizophrenia (phecode 295.1) and psychotic disorders (phecode 295) as a function of PRS. At each site, we estimated the odds ratios for both phecodes in the highest decile relative to the lowest, as well as to the remaining sample. Cross-site odds ratios were calculated by combining the log(OR) across healthcare systems through fixed-effect inverse variance-weighted meta-analysis using the metafor R package (https://cran.r-project.org/web/packages/metafor/). Finally, we calculated the median PRS percentile for cases and controls across all patients.

Schizophrenia PRS PheWAS. We conducted a PheWAS in each of the three healthcare systems using all phecodes with sufficient sample size. Logistic regressions between schizophrenia PRSs and each phecode were run with 10 ancestry principal components, median age across the EHR, sex, genotyping platform, and batch when available, included as covariates using the PheWAS R package[26]. 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 healthcare systems using a fixed-effect inverse variance-weighted model implemented using the PheWAS R package, which calls the meta R package (https://cran.r-project.org/web/packages/meta/). 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.

PheWAS Conditioning on Schizophrenia. To explore whether pleiotropic effects of the schizophrenia PRS were mediated by the diagnosis of schizophrenia itself, we also conducted PheWAS analyses using psychotic disorders (phecode 295; the broadest schizophrenia-related phecode) as an additional covariate.

Results

Our sample comprised of 91,980 patients (58% female) across three large healthcare systems that had collectively received over 10 million ICD-9 billing codes. At each healthcare site, the median electronic health record length was 9-13 years, included 50-110 unique visits, and over 100 ICD-9 codes (Table 1).

Table 1. Demographics and Clinical Characteristics

	GHS	PHS	VUMC
N	56,926	16,684	18,370
Mean age, years (SD)	58.6 (17.3)	57.7 (16.1)	60.7 (17.2)
Females, n (%)	35,111 (59%)	9,003 (54%)	9,350 (51%)
Total number of ICD-9 code days	4,958,475	3,856,860	2,474,372
Number of unique ICD-9 codes	11,433	10,347	13,801
Median number of visits per patient	110	52	50
Median EHR length, days	4,811	3,910	3,568
Median code days per patient	238	133	153

Mean age is the average age of the patient at their most recent hospital visit in which they received an ICD-9 code. A visit is both patient- and date-specific, but may include many individual ICD-9 codes. A code day is ICD-9-, patient-, and date-specific.

Penetrance of Schizophrenia PRS in Healthcare Systems

Overall case prevalence in our sample was 0.5% for schizophrenia (phecode 295.1) and 1.3% for schizophrenia and related psychotic disorders (phecode 295). Polygenic risk scores were robustly associated with schizophrenia in the cross-site meta-analysis (OR per standard deviation increase in PRS = 1.65 [95% confidence interval (CI), 1.5-1.8], $p = 1.25 \times 10^{-16}$) (Table 2); similar effects were observed in each individual healthcare system (Table S2). Patients in the highest risk decile had a four-fold increased odds of schizophrenia (phecode 295.1) compared to those in the bottom decile (95% CI, 2.4-6.5, $p = 4.43 \times 10^{-8}$) and a 2.1-fold increased odds of psychotic disorders (phecode 295) (95% CI, 1.6-2.7, $p = 6.40 \times 10^{-8}$; Figure 2). For the distribution of schizophrenia PRSs, the median schizophrenia case (phecode 295.1) was in the 66th PRS percentile, the median psychotic disorder case (phecode 295) was in the 59th PRS percentile, and the median control for both was in the 51st PRS percentile. Case prevalence in the top decile was 0.9% for schizophrenia (vs. 0.2% in the bottom decile) and 1.9% for psychosis (vs. 0.9% in the bottom decile). Those in the top decile also had a 2.1-fold increased odds of schizophrenia compared to those below the 90th percentile (95% CI, 1.6-2.6, $p = 4.69 \times 10^{-9}$) and a 1.5-fold increased odds of psychotic disorders (95% CI, 1.3-1.8, $p = 1.66 \times 10^{-7}$).

Schizophrenia PRS PheWAS

After excluding codes for which fewer than 100 cases were present at any site, we conducted PheWAS using 1338 disease categories across more than 90,000 patients. Phenome-wide significant associations with schizophrenia PRSs at each site were reported in Table S1. In total, the cross-site PheWAS meta-analysis yielded significant associations between schizophrenia PRSs and 26 medical phenotypes including schizophrenia (Table 2; Figure 1). As shown, the strongest 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[11]. We additionally found associations with two other psychiatric phenotypes, personality disorders and suicidal behavior. Many other syndromes were also significantly associated with schizophrenia PRS including obesity, heart palpitations, urinary syndromes and nonspecific somatic symptoms. Odds ratios for these phenotypes comparing the top versus bottom deciles ranged between 0.8 [95% CI, 0.7-0.8] for morbid obesity and 2.0 [95% CI, 1.5-2.7] for posttraumatic stress disorder (Figure 2). Case prevalence for all phenotypes across the PRS distribution was plotted in Figure S1.

PheWAS Conditioning on Schizophrenia

We explored whether some of the observed associations might be mediated through a clinical diagnosis of schizophrenia itself by conditioning the PheWAS on the broadest schizophrenia-related phecode (phecode 295). Nearly all associations remained significant and two novel associations met phenome-wide significance, diabetes mellitus and type 2 diabetes (Table S2; Figure S2). More specifically, associations remained significant with all five categories of anxiety disorders (anxiety disorder; anxiety, phobic and dissociative disorders; generalized anxiety disorder; posttraumatic stress disorder; and agoraphobia, social phobia, and panic disorder), all categories of mood disorders (mood disorders; bipolar disorder; depression; and major depressive disorder), 3 of the 4 categories of substance use disorders (substance addiction and disorders; tobacco use disorder; alcohol-related disorders), all obesity phenotypes (morbid obesity; obesity; bariatric surgery), malaise and fatigue, dysuria, and other urinary system symptoms. Associations with suicidal ideation, suicidal ideation or attempt, alcoholism, personality disorders, palpitations, and dysphagia no longer survived Bonferroni correction, although they remained top phenotypes (with *p*'s between 6.73×10^{-5} and 9.18×10^{-4}).

Table 2. Top Phenotypes from Schizophrenia PRS PheWAS Meta-Analysis

Phecode	Description	Beta	OR	SE	*p*	Total	Cases	Controls	Sites
300	Anxiety, phobic and dissociative disorders	0.098	1.103	0.009	4.20×10^{-26}	82,431	22,558	59,873	3
300.1	Anxiety disorder	0.101	1.106	0.010	1.62×10^{-25}	82,915	19,602	63,313	3
296.1	Bipolar	0.195	1.215	0.023	2.60×10^{-17}	90,529	2,529	88,000	3
295.1	Schizophrenia	0.499	1.647	0.060	1.25×10^{-16}	72,848	367	72,481	2
295	Schizophrenia and other psychotic disorders	0.267	1.306	0.033	4.02×10^{-16}	89,452	1,197	88,255	3
296	Mood disorders	0.072	1.074	0.009	7.33×10^{-16}	84,499	24,395	60,104	3
316	Substance addiction and disorders	0.168	1.183	0.022	1.24×10^{-14}	89,409	2,843	86,566	3
296.2	Depression	0.063	1.065	0.009	4.55×10^{-12}	84,544	22,683	61,861	3
318	Tobacco use disorder	0.068	1.070	0.011	1.30×10^{-10}	86,156	14,783	71,373	3
278.11	Morbid obesity	-0.068	0.934	0.011	1.47×10^{-10}	87,738	15,415	72,323	3
300.9	Posttraumatic stress disorder	0.208	1.231	0.036	8.93×10^{-9}	90,819	993	89,826	3
296.22	Major depressive disorder	0.085	1.088	0.015	2.34×10^{-8}	85,895	6,177	79,718	3
317	Alcohol-related disorders	0.113	1.120	0.020	3.09×10^{-8}	89,392	3,212	86,180	3
599.3	Dysuria	0.083	1.086	0.015	3.22×10^{-8}	82,015	6,684	75,331	3
300.11	Generalized anxiety disorder	0.093	1.098	0.017	3.95×10^{-8}	88,902	4,937	83,965	3
301	Personality disorders	0.213	1.237	0.041	1.52×10^{-7}	72,402	818	71,584	2
798	Malaise and fatigue	0.048	1.049	0.009	1.74×10^{-7}	74,628	22,033	52,595	3
599	Other symptoms/disorders or the urinary system	0.049	1.050	0.010	2.84×10^{-7}	78,757	20,594	58,163	3
297.1	Suicidal ideation	0.251	1.285	0.050	5.96×10^{-7}	72,236	534	71,702	2
297	Suicidal ideation or attempt	0.156	1.169	0.031	6.62×10^{-7}	89,321	1,359	87,962	3
278.1	Obesity	-0.044	0.957	0.009	6.88×10^{-7}	84,574	27,976	56,598	3
300.12	Agorophobia, social phobia, and panic disorder	0.115	1.122	0.025	2.80×10^{-6}	89,738	2,270	87,468	3
317.1	Alcoholism	0.110	1.116	0.024	3.55×10^{-6}	89,465	2,373	87,092	3
539	Bariatric surgery	-0.086	0.917	0.019	6.67×10^{-6}	90,784	3,818	86,966	3
532	Dysphagia	0.068	1.070	0.016	1.99×10^{-5}	86,191	5,438	80,753	3
427.9	Palpitations	0.058	1.059	0.014	2.08×10^{-5}	84,088	7,854	76,234	3

All effects listed surpassed the Bonferroni significance threshold (3.7×10^{-5}).

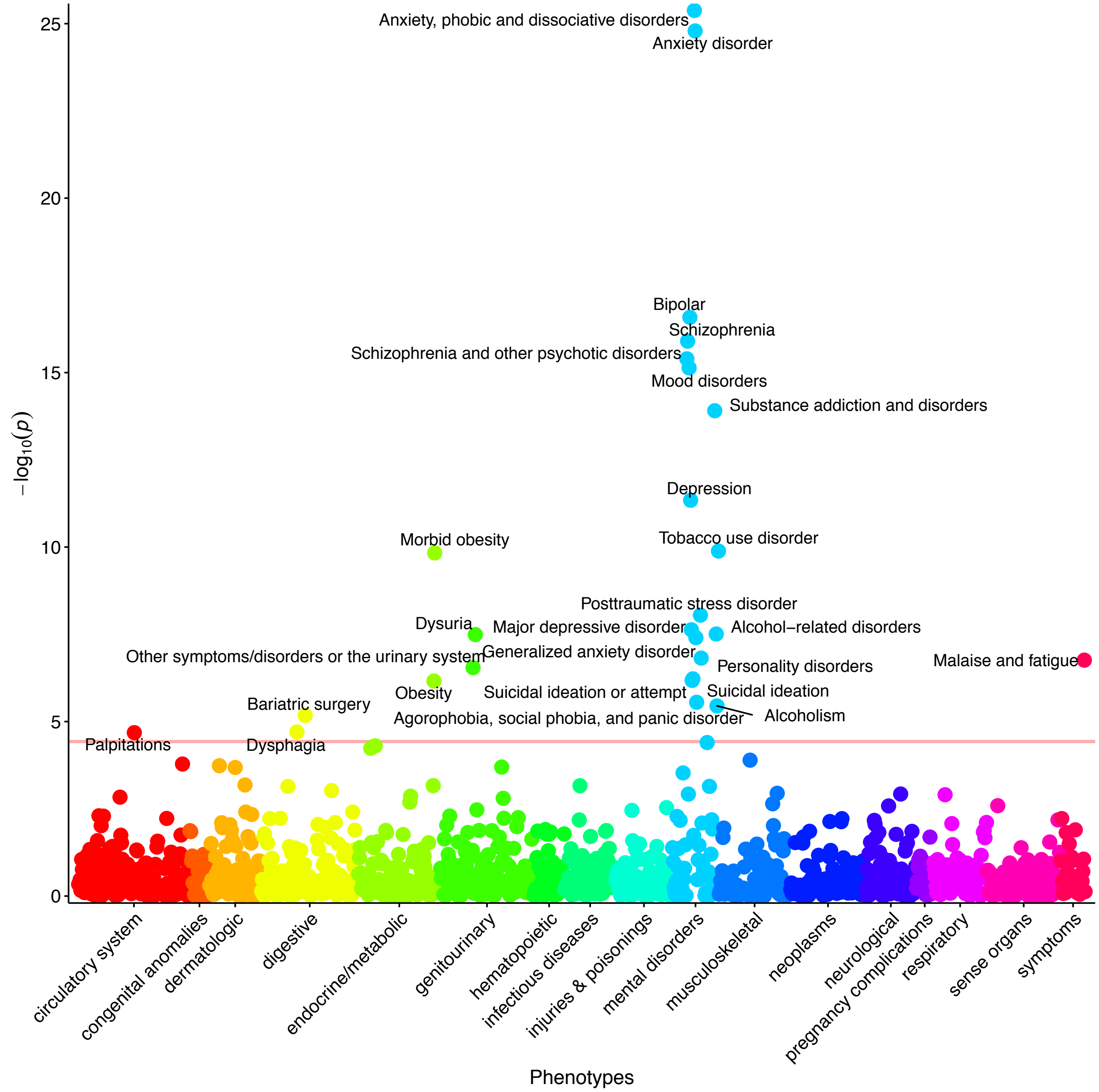

Figure 1. Schizophrenia PRS PheWAS Meta-Analysis. Manhattan plot for phenome-wide association with schizophrenia polygenic risk scores meta-analyzed across three healthcare systems (1338 phenotypes; 91,980 patients). The *x* axis is phenotype (grouped by broad disease category) and the *y* axis is significance ($-\log_{10} P$; 2-tailed) of association derived by logistic regression. The red line shows phenome-wide level significance (3.7 x 10^{-5}). 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 bariatric surgery.

Discussion

We investigated the impact of genetic risk for schizophrenia across the medical phenome in 91,980 patients from three, large healthcare systems. Several findings from our analysis are particularly noteworthy. First, in our cross-site meta-analysis, schizophrenia PRSs were highly statistically significantly associated with both “schizophrenia and related psychotic disorders” (*p*

= 4.02 x 10^{-16}) – a broad category of multiple psychosis-related ICD-9 codes – and schizophrenia itself (p = 1.25 x 10^{-16}), despite only 367 cases in the latter category. As expected, of the 1338 diagnostic categories examined, the largest effect size we observed was for schizophrenia. These results demonstrate that externally-derived polygenic risk scores can robustly detect risk for diagnosis of schizophrenia in healthcare settings using readily-available structured diagnostic codes.

At the same time, the effect size (an index of penetrance) was more modest than that reported in case-control cohorts ascertained for research purposes. For example, in the original report by the Psychiatric Genomics Consortium from which the risk scores were derived, individuals in the top decile of schizophrenia PRS relative to the bottom decile had a 7.8-20.3 increased odds of schizophrenia[16],whereas we observed an odds ratio of 4.0 (95% CI, 2.4 to 6.5). There are several potential reasons for this discrepancy, including differences in case and control definitions. Cases in the Psychiatric Genomics Consortium meta-analysis had to meet relatively stringent criteria based on clinical interviews by trained research personnel, and control ascertainment varied between studies, but often included screening for history of psychiatric or neurological disorders. This approach, typical for samples used for research, may maximize power for genetic discovery by extreme sampling from the tails of the genetic liability distribution. In contrast, our analysis was expressly designed to approximate use of a PRS in a typical clinical setting by applying a simple, easily-implementable definition for both cases (two or more schizophrenia-related ICD-9 codes) and controls (no schizophrenia-related codes). Thus, although the effect size we observed may have been attenuated due to some degree of misclassification, it may better reflect results that would be seen in real world clinical settings where PRS are applied to a broad healthcare population. In addition, we did not restrict the age range of cases and controls, which may have further reduced the apparent effect size of the schizophrenia PRS. This is because in our sample some fraction of high risk individuals who have not yet passed through the age of risk may have been misclassified as controls. Further research, including the application of natural language processing, may improve effect sizes by refining case and control definitions[28,29].

While the effect of the PRS we observed was not large enough on its own to stratify risk in a clinical setting (i.e., to discriminate between cases and controls on an individual level with high accuracy), it is comparable to those of non-genetic risk factors in established risk calculators. For example, two well-established coronary artery disease (CAD) risk factors – smoking and diabetes – were estimated in the Framingham Heart Study to have hazard ratios $\leq$ 2.0[30] – equivalent to the risk for the top schizophrenia PRS decile. Individuals in the top 5% of CAD PRS have risk of coronary disease (OR = 3.3 [95% CI, 3.1-3.6]) that is comparable to that seen among carriers of rare monogenic mutations causing hypercholesterolemia[15]. (Individuals in the top 5% of schizophrenia PRS had a 2.4 increased odds of schizophrenia, 95% CI, 1.8-3.3). The effects we observed were also similar to or greater than those seen in a recent PheWAS of PRSs for several common cancers (comparing top PRS quartile to bottom quartile; ORs = 1.3 - 3.3)[31]. Additionally, in a risk calculator for the transition to psychosis among high-risk individuals – one of the few individualized risk calculators developed within psychiatry – the best predictor was a symptom severity index with a hazard ratio of 2.1, 95% CI, 1.6-2.7[32].While this risk calculator was not validated for clinical use, it does reflect effects of variables used by clinicians to assess risk. In light of this, we speculate that incorporating genetic risk could be impactful within psychiatry. In future research, enhanced performance may be possible as the precision of PRSs increases (through larger sample size in discovery datasets)[14] and with

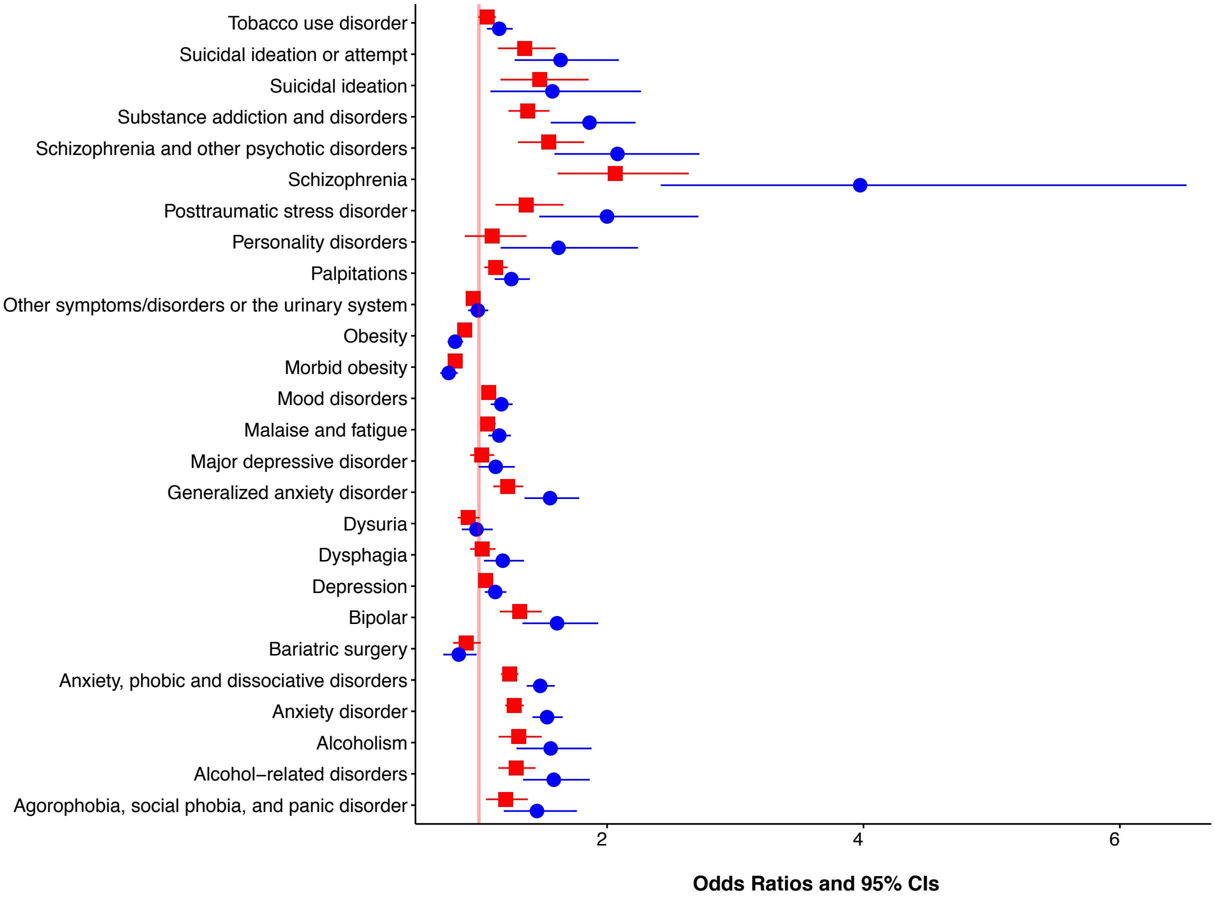

Figure 2. Odds ratios for top PRS decile. Odds ratios and 95% confidence intervals for phenotypes significant in meta-analysis were plotted for the top PRS decile with reference to both the remaining 90% (red squares) and the bottom decile (blue circles). The vertical red line reflects no change in risk (OR = 1).

refinement of EHR-based case definitions. It remains to be seen whether combining PRS risk estimates with other clinical predictors can meaningfully contribute to individualized risk assessment in psychiatry.

Besides increasing risk for schizophrenia, schizophrenia PRSs were associated with broader effects on mental health including increased risk for anxiety disorders, mood disorders, substance use disorders, personality disorders, and suicidal behavior. Anxiety disorders, mood disorders, and substance use disorders have all previously been linked to genetic risk for schizophrenia[11,33–35] and results reported here confirm in a clinical setting that these disorders share genetic risk. Certain personality disorders have also been previously linked to genetic liability for schizophrenia[36,37] (e.g., schizotypal or schizoid), although the phecode used here included all personality disorders. Personality disorders are common among patients with schizophrenia[38,39] and there is some evidence that personality dimensions in adolescence predict future psychopathology, including schizophrenia[40]. Nonetheless, our sensitivity analyses did not confirm any shared genetic liability. Similarly, suicidal ideation and attempt rates are much higher among patients with schizophrenia[41,42] and family history of schizophrenia has been associated with suicidal behavior[43]. However, our sensitivity analyses suggested that the relationship between suicidal behavior and schizophrenia in our sample may not be due to shared genetic risk.

Genetic liability for schizophrenia was associated with many non-psychiatric syndromes as well, including obesity and related phenotypes, urinary syndromes, palpitations, dysphagia, and malaise and fatigue. Most of these effects, while highly statistically significant, were not particularly large (ORs < 2). However, they may reveal interesting connections between schizophrenia genetic risk and other psychiatric and non-psychiatric diseases. For example, several obesity-related phenotypes, morbid obesity, obesity, and bariatric surgery were significantly negatively associated with schizophrenia PRSs (Table 2; Figure 1). Inverse phenotypic relationships between body mass index and schizophrenia have been observed previously in large studies[44–46]. Additionally, a recent report investigating bidirectional causal effects between these phenotypes (among others) found a significant genetic correlation using PRSs, but no evidence of causal effects, suggesting instead a shared genetic etiology[47]. Interestingly, there was also an inverse association between genetic liability for schizophrenia and diabetes, but only when conditioning on the diagnosis of schizophrenia. It may that this negative genetic correlation was masked in the primary analysis (i.e., when not conditioning on schizophrenia diagnosis) due to the opposing diabetogenic effects of antipsychotic medications[48].

We also found that dysuria – painful urination associated with a variety of causes including urinary tract infections (UTIs), sexually transmitted infections, yeast infections, and others – and a broader urinary system symptom category remained significantly associated with schizophrenia PRSs after conditioning on schizophrenia. In line with this, some epidemiological research has suggested rates of UTIs were higher in schizophrenia inpatients, relative to outpatients and controls[49], as well as in patients with non-affective psychosis relative to patients with psychotic and non-psychotic major depressive disorder[50]. However, the mechanism underlying this association is unclear; shared genetic risk is one possibility, but there could be other explanations as well that warrant further investigation. These 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, due to small numbers of patients of other ancestries, our analyses were restricted to patients of European descent, and the generalizability to other groups 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, ICD-10 diagnoses, setting of diagnosis (i.e., in- or outpatient; physician specialty), and treatment for the disease of interest. This was by design in order to mimic a real-world clinical population where PRSs may be implemented for clinical decision support. Future work incorporating these variables or expanding the case and control definitions to incorporate natural language processing algorithms may improve the predictive performance of PRSs and other risk factors for clinically-derived phenotypes. Third, specific associations within healthcare systems varied to some degree (Table S1), suggesting that results may vary according to the demographic and disease distributions in any given healthcare system. Relatedly, we noted disease prevalence was often lower in all patients in the healthcare system relative to the participants enrolled in the biobanks (a subset of those patients) across sites (Table S3). In general, case prevalence in the biobanks was more representative of population-level prevalence than was that in the healthcare systems[51], 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 healthcare system (Table S3). 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 PheWAS.

In conclusion, we demonstrate that an available measure of polygenic risk for schizophrenia is robustly associated with schizophrenia across three large healthcare systems using EHR data. While the observed penetrance of this schizophrenia PRS is attenuated in these settings compared to prior estimates derived from research cohorts, effect sizes are comparable to those seen for risk factors commonly used in clinical settings. We also find that polygenic risk for schizophrenia has pleiotropic effects on related psychiatric disorders as well as several non-psychiatric 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 healthcare systems.

Funding and Acknowledgements

We would like to acknowledge the Partners Biobank for providing samples, genomic data, and health information data. J.W.S. was supported in part by an NHGRI grant supporting the eMERGE Network (U01HG008685). J.W.S. is also a Tepper Family MGH Research Scholar and supported in part by the Demarest Lloyd, Jr. Foundation.

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, TN. The datasets used for the project were obtained from VUMCs 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, the 1S10RR025141-01 instrumentation award, and by the CTSA grant UL1TR000445 from NCATS/NIH. Additional funding provided by the NIH through grants P50GM115305 and U19HL065962. The authors wish to 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). Additionally, J.D. was supported by the Canadian Institutes of Health Research (MFE-142936), L.K.D. was supported by 1R01MH113362, and D.R. was supported by R01MH111776.

The authors would also like to 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 U01HG8679.

Conflicts of Interest

Dr. Smoller is an unpaid member of the Bipolar/Depression Research Community Advisory Panel of 23andMe. Dr. Kirchner received funding from Regeneron Genetics Center as part of the DiscovEHR study. All other authors have nothing to report.

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. doi:10.1001/jamapsychiatry.2014.2502.
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. doi:10.1016/S0140-6736(17)32154-2.
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. doi:10.1093/ije/dyu038.
4. Ahrnsbrak R, Bose J, Hedden SL, Lipari RN, Park-Lee E. Key substance use and mental health indicators in the United States: results from the 2016 national survey on drug use and health. *Subst Abus Ment Heal Serv Adm*. 2016. doi:10.1016/j.drugalcdep.2016.10.042.
5. Albert N, Melau M, Jensen H, Hastrup LH, Hjorthøj C, Nordentoft M. The effect of duration of untreated psychosis and treatment delay on the outcomes of prolonged early intervention in psychotic disorders. *npj Schizophr*. 2017. doi:10.1038/s41537-017-0034-4.
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. doi:10.1016/j.schres.2014.01.022.
7. Amminger GP, Edwards J, Brewer WJ, Harrigan S, McGorry PD. Duration of untreated psychosis and cognitive deterioration in first-episode schizophrenia. *Schizophr Res*. 2002. doi:10.1016/S0920-9964(01)00278-X.
8. Habert J, Katzman MA, Oluboka OJ, et al. Functional Recovery in Major Depressive Disorder: Focus on Early Optimized Treatment. *Prim Care Companion CNS Disord*. 2016. doi:10.4088/PCC.15r01926.
9. Kvitland LR, Ringen PA, Aminoff SR, et al. Duration of untreated illness in first-treatment bipolar I disorder in relation to clinical outcome and cannabis use. *Psychiatry Res*. 2016. doi:10.1016/j.psychres.2016.07.064.
10. Wang PS, Berglund P, Olfson M, Pincus HA, Wells KB, Kessler RC. Failure and delay in initial treatment contact after first onset of mental disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry*. 2005. doi:10.1001/archpsyc.62.6.603.
11. Brainstorm Consortium TB, Anttila V, Bulik-Sullivan B, et al. Analysis of shared heritability in common disorders of the brain. *Science*. 2018;360(6395):eaap8757. doi:10.1126/science.aap8757.
12. Smoller JW, Andreassen OA, Edenberg HJ, Faraone S V., Glatt SJ, Kendler KS. Psychiatric genetics and the structure of psychopathology. *Mol Psychiatry*. January 2018:1. doi:10.1038/s41380-017-0010-4.
13. 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. doi:10.1016/j.biopsych.2016.06.028.
14. Zheutlin AB, Ross DA. Polygenic Risk Scores: What Are They Good For? *Biol Psychiatry*. 2018. doi:10.1016/j.biopsych.2018.04.007.
15. Khera A V., Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*.

August 2018:1. doi:10.1038/s41588-018-0183-z.

16. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014;511(7510):421-427. doi:10.1038/nature13595.
17. Crawford DC, Crosslin DR, Tromp G, et al. EMERGEing progress in genomics-the first seven years. *Front Genet*. 2014. doi:10.3389/fgene.2014.00184.
18. Gottesman O, Kuivaniemi H, Tromp G, et al. The Electronic Medical Records and Genomics (eMERGE) Network: Past, present, and future. *Genet Med*. 2013. doi:10.1038/gim.2013.72.
19. Smoller JW. The use of electronic health records for psychiatric phenotyping and genomics. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*. 2017.
20. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: Challenges and strategies. *Nat Rev Genet*. 2013. doi:10.1038/nrg3461.
21. Danciu I, Cowan JD, Basford M, et al. Secondary use of clinical data: The Vanderbilt approach. *J Biomed Inform*. 2014. doi:10.1016/j.jbi.2014.02.003.
22. 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. doi:10.1038/gim.2015.187.
23. Karlson EW, Boutin NT, Hoffnagle AG, Allen NL. Building the partners healthcare biobank at partners personalized medicine: Informed consent, return of research results, recruitment lessons and operational considerations. *J Pers Med*. 2016. doi:10.3390/jpm6010002.
24. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. *Bioinformatics*. 2015. doi:10.1093/bioinformatics/btu848.
25. 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. doi:10.1371/journal.pone.0175508.
26. Carroll RJ, Bastarache L, Denny JC. R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*. 2014. doi:10.1093/bioinformatics/btu197.
27. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Informatics Assoc*. 2016. doi:10.1093/jamia/ocv130.
28. Perlis RH, Iosifescu D V., Castro VM, et al. Using electronic medical records to enable large-scale studies in psychiatry: Treatment resistant depression as a model. *Psychol Med*. 2012. doi:10.1017/S0033291711000997.
29. Castro VM, Minnier J, Murphy SN, et al. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *Am J Psychiatry*. 2015. doi:10.1176/appi.ajp.2014.14030423.
30. D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation*. 2008. doi:10.1161/CIRCULATIONAHA.107.699579.
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. *American Journal of Human Genetics*. 2018.

32. Cannon TD, Yu C, Addington J, et al. An individualized risk calculator for research in prodromal psychosis. *Am J Psychiatry*. 2016;173(10):980-988. doi:10.1176/appi.ajp.2016.15070890.

33. Gandal MJ, Haney JR, Parikshak NN, et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science*. 2018;359(6376):693-697. doi:10.1126/science.aad6469.

34. 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(9875):1371-1379. doi:10.1016/S0140-6736(12)62129-1.

35. Mistry S, Harrison JR, Smith DJ, Escott-Price V, Zammit S. The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: Systematic review. *Schizophr Res*. 2017. doi:10.1016/j.schres.2017.10.037.

36. Nelson MT, Seal ML, Pantelis C, Phillips LJ. Evidence of a dimensional relationship between schizotypy and schizophrenia: A systematic review. *Neurosci Biobehav Rev*. 2013. doi:10.1016/j.neubiorev.2013.01.004.

37. Bigdeli TB, Bacanu SA, Webb BT, et al. Molecular validation of the schizophrenia spectrum. *Schizophr Bull*. 2014. doi:10.1093/schbul/sbt122.

38. Newton-Howes G, Tyrer P, North B, Yang M. The prevalence of personality disorder in schizophrenia and psychotic disorders: Systematic review of rates and explanatory modelling. *Psychol Med*. 2008. doi:10.1017/S0033291707002036.

39. Wei Y, Zhang T, Chow A, et al. Co-morbidity of personality disorder in schizophrenia among psychiatric outpatients in China: data from epidemiologic survey in a clinical population. *BMC Psychiatry*. 2016. doi:10.1186/s12888-016-0920-8.

40. 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. doi:10.1177/0004867415569796.

41. Pedersen CG, Jensen SOW, Gradus J, Johnsen SP, Mainz J. Systematic suicide risk assessment for patients with schizophrenia: a national population-based study. *Psychiatr Serv*. 2014. doi:10.1176/appi.ps.201200021.

42. Dutta R, Murray RM, Hotopf M, Allardyce J, Jones PB, Boydell J. Reassessing the long-term risk of suicide after a first episode of psychosis. *Arch Gen Psychiatry*. 2010. doi:10.1001/archgenpsychiatry.2010.157.

43. 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. doi:10.1016/j.schres.2016.12.001.

44. Zammit S, Rasmussen F, Farahmand B, et al. Height and body mass index in young adulthood and risk of schizophrenia: A longitudinal study of 1 347 520 Swedish men. *Acta Psychiatr Scand*. 2007. doi:10.1111/j.1600-0447.2007.01063.x.

45. Sørensen HJ, Mortensen EL, Reinisch JM, Mednick SA. Height, weight and body mass index in early adulthood and risk of schizophrenia. *Acta Psychiatr Scand*. 2006. doi:10.1111/j.1600-0447.2006.00784.x.

46. Duncan LE, Shen H, Ballon JS, Hardy K V, Noordsy DL, Levinson DF. Genetic Correlation Profile of Schizophrenia Mirrors Epidemiological Results and Suggests Link Between Polygenic and Rare Variant (22q11.2) Cases of Schizophrenia. *Schizophr Bull*. 2017. doi:10.1093/schbul/sbx174.

47. So H-C, Chau K-L, Ao F-K, Mo C-H, Sham P-C. Exploring shared genetic bases and

causal relationships of schizophrenia and bipolar disorder with 28 cardiovascular and metabolic traits. *Psychol Med*. 2018. doi:10.1017/S0033291718001812.

48. Annamalai A, Kosir U, Tek C. Prevalence of obesity and diabetes in patients with schizophrenia. *World J Diabetes*. 2017. doi:10.4239/wjd.v8.i8.390.

49. Miller BJ, Graham KL, Bodenheimer CM, Culpepper NH, Waller JL, Buckley PF. A prevalence study of urinary tract infections in acute relapse of schizophrenia. *J Clin Psychiatry*. 2013. doi:10.4088/JCP.12m08050.

50. Carson CM, Phillip N, Miller BJ. Urinary tract infections in children and adolescents with acute psychosis. *Schizophr Res*. 2017. doi:10.1016/j.schres.2016.11.004.

51. Kessler RC, Chiu WT, Demler O, Merikangas KR, Walters EE. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry*. 2005. doi:10.1001/archpsyc.62.6.617.