

## ARTICLE OPEN



## Developing a phenotype risk score for tic disorders in a large, clinical biobank

Tyne W. Miller-Fleming<sup>1,2</sup> , Annmarie Allos<sup>1,3</sup>, Emily Gantz<sup>4,5,6</sup> , Dongmei Yu<sup>7,8</sup>, David A. Isaacs<sup>5,6</sup> , Carol A. Mathews<sup>9</sup> , Jeremiah M. Scharf<sup>7,8</sup> and Lea K. Davis<sup>1,2,10,11,12</sup>

© The Author(s) 2024

Tics are a common feature of early-onset neurodevelopmental disorders, characterized by involuntary and repetitive movements or sounds. Despite affecting up to 2% of children and having a genetic contribution, the underlying causes remain poorly understood. In this study, we leverage dense phenotype information to identify features (i.e., symptoms and comorbid diagnoses) of tic disorders within the context of a clinical biobank. Using de-identified electronic health records (EHRs), we identified individuals with tic disorder diagnosis codes. We performed a phenome-wide association study (PheWAS) to identify the EHR features enriched in tic cases versus controls ( $n = 1406$  and  $7030$ ; respectively) and found highly comorbid neuropsychiatric phenotypes, including: obsessive-compulsive disorder, attention-deficit/hyperactivity disorder, autism spectrum disorder, and anxiety ( $p < 7.396 \times 10^{-5}$ ). These features (among others) were then used to generate a phenotype risk score (PheRS) for tic disorder, which was applied across an independent set of 90,051 individuals. A gold standard set of tic disorder cases identified by an EHR algorithm and confirmed by clinician chart review was then used to validate the tic disorder PheRS; the tic disorder PheRS was significantly higher among clinician-validated tic cases versus non-cases ( $p = 4.787 \times 10^{-151}$ ;  $\beta = 1.68$ ;  $SE = 0.06$ ). Our findings provide support for the use of large-scale medical databases to better understand phenotypically complex and underdiagnosed conditions, such as tic disorders.

*Translational Psychiatry* (2024)14:311; <https://doi.org/10.1038/s41398-024-03011-w>

## INTRODUCTION

Tic disorders (TD) are the most common movement disorder in children and are characterized by sudden and recurrent movements and/or vocalizations [1–4]. While many tic symptoms resolve within a year, persistent TD can cause disruptions to daily life and may have long-term effects on an individual's social, physical, and mental health [5–7]. TD is highly comorbid with several other psychiatric and neurodevelopmental conditions, including obsessive-compulsive disorder (OCD), attention-deficit/hyperactivity disorder (ADHD), and autism spectrum disorder (ASD), among others [8–17]. One study found that 86% of individuals diagnosed with the most common tic disorder, Tourette syndrome, are diagnosed with at least one additional psychiatric disorder during their lifetime, and up to 58% of Tourette syndrome patients are diagnosed with two or more additional psychiatric disorders during their lifetime [13]. This phenotypic heterogeneity complicates the diagnosis and treatment of patients with Tourette syndrome and other TDs.

Tic disorders are both phenotypically and genetically complex. Heritability measurements for Tourette syndrome range from 0.58 to 0.77 and suggest a strong underlying genetic component;

however, identifying the genetic signatures of TD has been difficult [18–20]. Familial and patient studies have identified candidate genes, none of which have been confirmed in independent investigations [21–29]. Genome-wide association studies of TD and Tourette syndrome demonstrate high polygenicity and have identified a few genome-wide significant signals, including Collagen Type XXVII Alpha 1 chain, *COL27A1* and Fms Related Receptor Tyrosine Kinase 3, *FLT-3*, though neither locus has been replicated in an independent study to date [30–32]. This lack of replicable signals is likely due to the polygenic nature of TD, in addition to the challenge of recruiting large cohorts of TD patients [33].

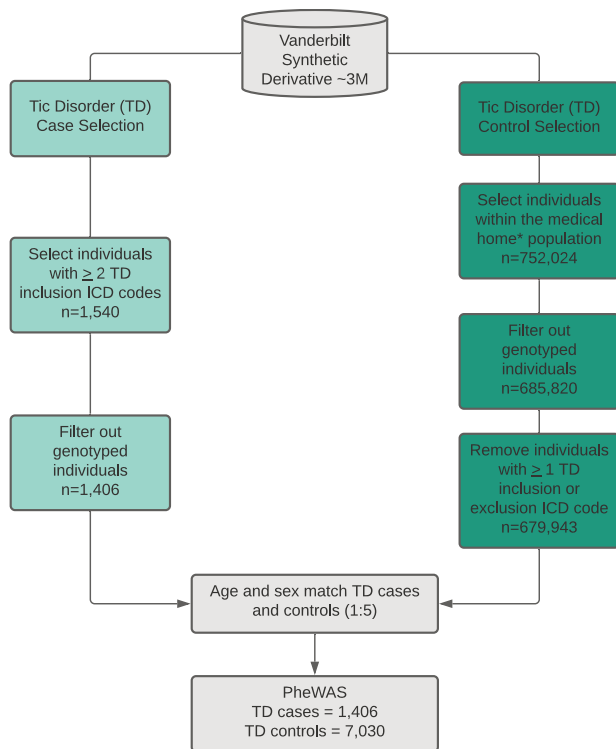
Electronic health records (EHRs) are a useful resource for studying disease outcomes and comorbidities [34]. EHR systems often date back decades and document a wide range of phenotype information, including diagnosis and billing codes, clinician notes, medical histories, lab results, medications, and procedural codes. Additionally, EHR systems allow for the investigation of individuals across diverse disease groups without requiring the resources needed to recruit large cohorts of individuals for genomic studies. In the case of phenotypically

<sup>1</sup>Vanderbilt Genetics Institute, Vanderbilt University Medical Center, TN, Nashville, USA. <sup>2</sup>Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. <sup>3</sup>Department of Cognitive Science, Dartmouth College, Hanover, NH, USA. <sup>4</sup>Department of Pediatric Neurology, Children's Hospital of Alabama, Birmingham, AL, USA. <sup>5</sup>Department of Neurology, Vanderbilt University Medical Center, Nashville, TN, USA. <sup>6</sup>Department of Pediatrics, Monroe Carell Jr. Children's Hospital at Vanderbilt, Nashville, TN, USA. <sup>7</sup>Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. <sup>8</sup>Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>9</sup>Department of Psychiatry, Genetics Institute, Center for OCD, Anxiety and Related Disorders, University of Florida, Gainesville, FL, USA. <sup>10</sup>Department of Biomedical Informatics, Vanderbilt University Medical Center, TN, Nashville, USA. <sup>11</sup>Department of Psychiatry and Behavioral Sciences, Vanderbilt University Medical Center, TN, Nashville, USA. <sup>12</sup>Department of Molecular Physiology and Biophysics, Vanderbilt University, TN, Nashville, USA.

✉email: [tyne.w.miller-fleming@vumc.org](mailto:tyne.w.miller-fleming@vumc.org); [lea.k.davis@vumc.org](mailto:lea.k.davis@vumc.org)

Received: 17 March 2023 Revised: 28 June 2024 Accepted: 4 July 2024

Published online: 28 July 2024



**Fig. 1 Extraction of tic disorder (TD) cases and controls from electronic health records (EHR).** TD cases were identified by selecting individuals with 2 instances of the TD inclusion phenotypes. TD cases were restricted to the non-genotyped population in the EHR. TD controls were matched to TD cases after filtering out the genotyped individuals, those that did not belong to the medical home, and those with at least 1 inclusion or exclusion ICD code for TD. Medical home individuals are those that have visited a Vanderbilt clinic at least 5 times over a three-year period. Matching of cases and controls was performed at a 1:5 ratio, respectively, based on current age and sex. ICD9/10 billing codes used as inclusion/exclusion criteria for TD cases and controls are listed in Supplemental Table 1.

complex conditions such as TD, EHRs can provide dense phenotype information spanning before and after diagnosis.

The recent use of phenotype risk scores (PheRS) calculated from medical records has successfully identified patients that exhibit overlapping features of disease. Similar to genetic risk scores, a discovery cohort is used to identify the phenotypic features that characterize a disease or condition in the medical records. These features are then evaluated within an independent target population, and each individual is assigned a score based on the number of features they exhibit. Individuals with a high PheRS are phenotypically similar to the diagnosed individuals of the discovery dataset, whereas individuals with a low PheRS share little or no “phenome” (i.e., set of all phenotypes expressed) with diagnosed individuals. The PheRS method was initially developed to identify patients with features of Mendelian diseases within the medical record database but has recently been applied to common neuropsychiatric conditions, including major depressive disorder, generalized anxiety disorder, and posttraumatic stress disorder [35–38]. This method condenses the medical phenome into a single quantitative score, which can then be used for downstream analyses, making this tool useful for the evaluation of phenotypically complex conditions. Additionally, because this method relies solely on the collection of diagnosis codes within the medical record, it is a powerful tool for indexing liability for TD, given that tics are a common comorbidity but may not be explicitly coded or charted in a medical record. Similar EHR-based

machine-learning methods have been successful in the identification of patients with underdiagnosed phenotypes, such as developmental stuttering [39].

In this study we leverage de-identified medical records for 3.6 million individuals from the Vanderbilt EHR system to identify the phenotypic correlates of TD. Using diagnosis billing codes, we identified 1,406 individuals with TD diagnoses and 7,030 age and sex-matched controls. A phenome-wide association study identified 69 phenotypes that were significantly associated with TD diagnosis, including several psychiatric and neurological phenotypes. Using the results of the PheWAS, we generated a PheRS and deployed it within an independent cohort of 90,051 individuals, including 266 individuals with clinician-validated TD diagnoses. We found that the TD PheRS was significantly higher for clinically validated tic patients versus non-diagnosed individuals. This proof-of-concept study: (1) supports the utility of medical records for evaluating the longitudinal effects of phenotypically complex diseases such as TD and (2) provides a framework for using the PheRS as a tool for phenome-wide investigations, improving sample sizes for downstream genetic analyses.

## RESULTS

### PheWAS identified the complex clinical phenome of TD patients

To uncover the phenotypic correlates of TD within a database of de-identified EHR, we performed a PheWAS of TD diagnosis presence/absence (Fig. 1, Supplemental Table 1). As expected for an early-onset disorder with a male-bias, our EHR-derived TD cases (and matched controls) were relatively young (average age 24.22/24.25 years old, average median age of medical record 15.40/13.95 years old, and average age at first ICD code 11.39/9.02 years old in cases/controls) and predominantly male (72.48%/72.59% male in cases/controls). Consistent with the demographic characteristics of the EHR at Vanderbilt, TD cases and controls were significantly skewed for individuals with EHR-reported race and ethnicity as white and non-Hispanic (77.67%/75.02% white in cases/controls and 84.57%/89.20% non-Hispanic in cases/controls) (Table 1). PheWAS analyses stratified by EHR-reported race were difficult to interpret because of the small sample sizes among the non-white populations (Supplemental Table 2), so we performed our PheWAS across all patients and used the EHR-reported race and ethnicity variables as covariates in the PheWAS model. We identified 69 phenotypes significantly associated with TD case status after Bonferroni-correction ( $p < 7.396 \times 10^{-5}$ ; Fig. 2 and Supplemental Table 2). These phenotypes represent clinical diagnoses, which have been previously mapped from ICD9/10 billing codes extracted from the EHR [40]. Included among the top associations were tic disorders and tics of organic origin, both of which were significantly associated with the TD case label ( $p = 4.18 \times 10^{-65}$ ;  $\beta = 9.94$ ; SE = 0.58 and  $p = 1.40 \times 10^{-42}$ ;  $\beta = 6.92$ ; SE = 0.51, respectively).

### Psychiatric and neurological conditions were highly comorbid in TD patients

As expected, in addition to tic-specific phenotypes, the top signals from the TD PheWAS included known comorbidities, such as: anxiety disorders ( $p = 1.12 \times 10^{-152}$ ;  $\beta = 2.38$ ; SE = 0.09) [13], attention-deficit/hyperactivity disorder ( $p = 1.84 \times 10^{-150}$ ;  $\beta = 3.43$ ; SE = 0.13) [12], autism ( $p = 1.17 \times 10^{-99}$ ;  $\beta = 3.57$ ; SE = 0.17) [15], developmental delays and disorders ( $p = 3.99 \times 10^{-76}$ ;  $\beta = 2.98$ ; SE = 0.16), conduct disorders ( $p = 2.57 \times 10^{-72}$ ;  $\beta = 3.41$ ; SE = 0.19) [8], mood disorders ( $p = 2.57 \times 10^{-68}$ ;  $\beta = 1.83$ ; SE = 0.10), obsessive-compulsive disorders ( $p = 1.86 \times 10^{-52}$ ;  $\beta = 4.29$ ; SE = 0.28) [16], depression ( $p = 9.46 \times 10^{-44}$ ;  $\beta = 1.73$ ; SE = 0.13) [41], migraine ( $p = 2.70 \times 10^{-35}$ ;  $\beta = 1.73$ ; SE = 0.14) [42], and sleep disorders ( $p = 1.19 \times 10^{-31}$ ;  $\beta = 1.39$ ; SE = 0.12) [43, 44]. The most significantly associated phenotype was pediatric

**Table 1.** Demographics for Tic Disorder PheWAS. Tic Disorder (TD) case-control demographics.

	Cases	Controls	p value
N	1406	7030	–
Sex (%male)	72.48	72.59	0.93
EHR-reported ethnicity (%Hispanic, %Non-Hispanic, %Unknown)	2.63, 84.57, 12.80	7.14, 89.20, 3.66	2.2e-16 <sup>b</sup>
EHR-reported race <sup>a</sup> (%A, %B, %D, %I, %N, %O, %P, %U, %W)	1.3, 6.8, 0.28, 0.14 0.5, 0.07, 0, 13.58, 77.67	1.96, 17.4, 0.27, 0.17, 0.92, 0.03, 0, 4.35, 75.02	2.2e-16 <sup>b</sup>
Current age (mean+SD)	24.22 + 14.03	24.25 + 14.01	0.95
Median age of record (mean+SD)	15.40 + 13.30	13.95 + 14.20	2.38e-4 <sup>b</sup>
Age at first ICD code (mean+SD)	11.39 + 13.20	9.02 + 13.60	1.18e-09 <sup>b</sup>
Age at last ICD code (mean+SD)	18.12 + 13.47	18.06 + 13.89	0.87
Visits (mean+SD)	29.20 + 53.92	25.18 + 37.83	7.86e-3 <sup>b</sup>
Medical record length (mean+SD)	6.74 + 5.70	9.04 + 4.43	3.94e-44 <sup>b</sup>
ICD code density (mean+SD)	83.21 + 202.04	65.64 + 130.0	0.002 <sup>b</sup>
Mean codes per visit (mean+SD)	2.76 + 1.63	2.39 + 1.13	1.50e-15 <sup>b</sup>
Median codes per visit (mean+SD)	2.28 + 1.54	1.87 + 0.98	1.59e-20 <sup>b</sup>

Tic Disorder (TD) case-control demographics. TD cases ( $n = 1406$ ) were identified using billing codes within the electronic health records. Tic disorder controls ( $n = 7030$ ) were matched to cases based on current age and sex. Independent-samples  $t$ -test or  $\chi^2$  analyses were performed on all measures to determine if there was a statistically significant difference between cases and controls.

<sup>a</sup>EHR-reported race abbreviations: (A): Asian, (B): Black, (D): Declined, (I): Alaskan/Indian, (N): Other Race, (O): Not listed, (P): Pacific Island, (U): Unknown, (W): White.

<sup>b</sup>Denotes statistically significant  $p$ -value.

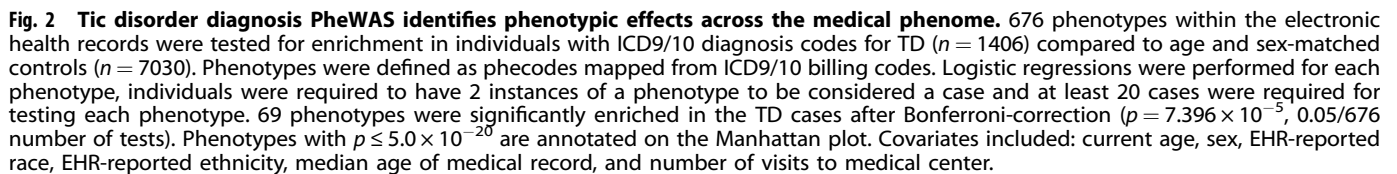
neuropsychiatric disorders ( $p < 5 \times 10^{-324}$ ;  $\beta = 4.80$ ; SE, 0.11), a phecode label that encompasses multiple disorders of impaired communication and socialization skills, including autistic disorders, Asperger's syndrome, Rett's disorder, and childhood disintegrative disorder. In the EHR, the phecodes for attention-deficit/hyperactivity disorder (313.1), tic disorders (313.2), and ASD (313.3) are also subcategories within the *pediatric neuropsychiatric disorders* phenotype (313), which explains our finding that 83.36% of the EHR-defined TD cases also had a diagnosis for pediatric neuropsychiatric disorders (Supplemental Table 3).

The TD case label was enriched for several other neuropsychiatric phenotypes, such as bipolar disorder ( $p = 5.11 \times 10^{-41}$ ;  $\beta = 2.57$ ; SE = 0.19), suicidal ideation ( $p = 1.38 \times 10^{-25}$ ;  $\beta = 2.02$ ; SE = 0.19), schizophrenia and other psychotic disorders ( $p = 1.16 \times 10^{-24}$ ;  $\beta = 2.61$ ; SE = 0.25), transient alteration of awareness ( $p = 6.33 \times 10^{-20}$ ;  $\beta = 2.16$ ; SE = 0.24), adjustment reaction ( $p = 1.52 \times 10^{-17}$ ;  $\beta = 1.46$ ; SE = 0.17), personality disorders ( $p = 1.89 \times 10^{-13}$ ;  $\beta = 2.90$ ; SE = 0.39), posttraumatic stress disorder ( $p = 6.93 \times 10^{-8}$ ;  $\beta = 1.32$ ; SE = 0.24), and eating disorder ( $p = 2.37 \times 10^{-5}$ ;  $\beta = 1.59$ ; SE = 0.38) (Supplemental Table 3). Among EHR-derived TD cases, 85.6% had at least one additional neuropsychiatric disorder diagnosis, and 51.9% had two or more additional neuropsychiatric diagnoses, compared to 21.2% and 16.8%, respectively, within controls. These differences are statistically significant ( $p < 1 \times 10^{-4}$ , Fisher's Exact Test) and are comparable to previous findings of psychiatric comorbidities within individuals with TD [13]. The proportion of control individuals with at least one neuropsychiatric diagnosis was higher than expected (7%), likely because of ascertainment bias of young individuals within the EHR, which tends to be enriched for neurodevelopmental phenotypes. Nevertheless, we found that 47% of the TD cases received an ICD code for psychiatric disorders or neurological conditions at their first medical center visit, compared to 7% of the controls (Supplemental Table 4). Because our EHR database contains longitudinal phenotype information, we examined whether tic diagnoses were more commonly made prior to or after other psychiatric diagnoses. Greater than 64% of

the TD cases were diagnosed with a tic disorder before receiving any additional neuropsychiatric diagnoses (Supplemental Table 5). These findings are consistent with the early onset nature of tic disorders and may provide clinicians with critical information regarding the comorbidities that TD patients are at greatest risk of developing later in life.

The neurological phenotypes enriched in TD cases (extrapyramidal disease and abnormal movement disorders ( $p = 1.11 \times 10^{-135}$ ;  $\beta = 5.10$ ; SE = 0.21), abnormal movement ( $p = 4.58 \times 10^{-62}$ ;  $\beta = 2.00$ ; SE = 0.12), and torsion dystonia ( $p = 6.82 \times 10^{-20}$ ;  $\beta = 3.43$ ; SE = 0.38)) may represent comorbidity, misdiagnoses, or the evolution of diagnoses during a diagnostic odyssey. Abnormal movements can also occur in tandem with TD, as a secondary response to TD medication, or independently of TD [45]. To further investigate this, we performed a sensitivity analysis, conditioning the TD PheWAS on the presence or absence of commonly prescribed medications for tics. Associations between hyperkinetic movement disorders and tic disorder diagnosis status remained after correcting for medications, suggesting that these movement phenotype associations were not solely side effects of TD medication usage (Supplemental Table 6) [46, 47]. We found that within TD cases, 54% of individuals with a hyperkinetic movement diagnosis received a tic disorder diagnosis first, while the remaining 46% received a movement diagnosis first; however, it is difficult to disentangle these findings, which cannot rule out multiple possibilities for the co-occurring codes, including clinical misdiagnoses (Supplemental Table 7) [45].

The 676 phenotypes assessed in our TD PheWAS were binned across 17 distinct phenotype groups, such as psychiatric disorders, neurological phenotypes, neoplasms, etc. To determine whether a specific phenotype group was over-represented among the 69 significant results, we used a hypergeometric test. Over 80% of the phenotypes enriched in TD cases mapped to psychiatric and neurological disorders (7.19-fold enrichment,  $p = 9.318 \times 10^{-30}$  and 4.66-fold enrichment,  $p = 8.745 \times 10^{-11}$ , respectively) (Fig. 3A, B). The remaining significantly enriched

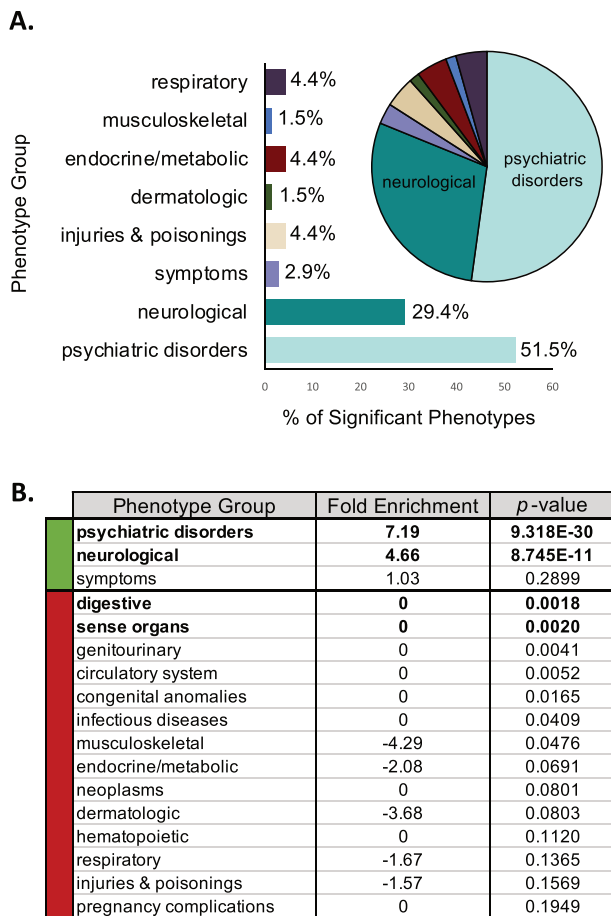


reveal the complex phenotypic architecture within individuals diagnosed with TD.

To consolidate the complex medical phenome of TD into a single, quantitative measure, we constructed a phenotype risk score (PheRS) using the 69 phenotypic features identified by the TD PheWAS (Supplemental Table 2 and Fig. 4A). We identified the PheRS features in the non-genotyped population of the Vanderbilt SD (discovery sample) by performing the PheWAS described above. The significantly associated phenotypes were then used to calculate the PheRS in the genotyped BioVU population (target sample) (Supplemental Tables 8 and 9). As expected, most patients in our target population (59.83%) had a very low TD PheRS because they did not exhibit any phenotypic features of TD (Fig. 4B).

Across the target population, we found that individuals of European ancestry (EA) had a significantly higher mean TD PheRS compared to individuals of African ancestry (AA); however, the EA sample size was much larger which may contribute to this finding ( $p < 2 \times 10^{-16}$ ;  $\beta = 0.1$ ;  $SE = 8.75 \times 10^{-3}$ ,  $n_{EA/AA} = 70,439/15,174$ ). This finding may also reflect a difference in healthcare utilization between individuals of European and African ancestry, as we found that although these populations had similar medical record lengths, individuals of African ancestry had significantly fewer medical center visits compared to individuals of European ancestry ( $p = 4.72 \times 10^{-9}$ ;  $\beta = 5.38 \times 10^{-4}$ ;  $SE = 9.17 \times 10^{-5}$ ). We further found that the TD





**Fig. 3 Tic disorder PheWAS is enriched for psychiatric disorders and neurological phenotypes.** **A** 69 phenotypes were significantly enriched or depleted within the tic disorder cases versus controls. The bar graph and pie chart show the proportions of these 69 phenotypes across 17 phenotype groups. Over 80% of the phenotypes enriched in the tic disorder individuals belong to the psychiatric disorders or neurological phenotype groups. **B** The phenotype groups for psychiatric disorders and neurological phenotypes are significantly enriched in the tic disorder individuals (7.19- and 4.66-fold enrichment, respectively,  $p < 0.00294$  calculated by the hypergeometric test). Digestive disorders and phenotypes within the sense organs were significantly depleted within the TD cases compared to controls ( $p < 0.00294$  calculated by the hypergeometric test). Descriptions of the pcode/ICD codes that map to each category have been previously described and can be found here (<https://phewascatalog.org/>).

PheRS was significantly higher in females compared to males across the BioVU population ( $p = 6.02 \times 10^{-16}$ ;  $\beta = -0.05$ ;  $SE = 6.6 \times 10^{-3}$ ,  $n_{\text{female/male}} = 51,183/38,865$ ). Although there is a well-documented male bias among TD, there is literature to suggest that the male bias within TD attenuates with age and that females with TD have more severe and persistent symptoms, consistent with our findings [48–52]. It is also possible that this difference reflects a bias in healthcare utilization, as female patients in BioVU had significantly more clinic visits ( $p = 0.002$ ,  $\beta = -2.18$ ,  $SE = 0.69$ ) and significantly longer records ( $p < 2 \times 10^{-16}$ ;  $\beta = -1.17$ ,  $SE = 0.05$ ) compared to male BioVU patients. The TD PheRS was also inversely associated with current age ( $p < 2 \times 10^{-16}$ ;  $\beta = -3.91 \times 10^{-3}$ ,  $SE = 1.45 \times 10^{-4}$ ) and median age of medical record ( $p < 2 \times 10^{-16}$ ;  $\beta = -3.46 \times 10^{-3}$ ,  $SE = 1.46 \times 10^{-4}$ ), suggesting that younger individuals are more likely to receive codes that contribute to the TD PheRS. Despite being younger, we found that individuals with a higher TD PheRS also had a greater number of visits to the medical center ( $p < 2 \times 10^{-16}$ ,

$\beta = 5.38 \times 10^{-3}$ ;  $SE = 3.98 \times 10^{-5}$ ) (Supplemental Table 10). This finding was expected, as the TD PheRS increases with accumulation of relevant phecodes during visits to the medical center.

### TD PheRS was significantly higher in clinically validated tic cases

We previously generated and deployed an algorithm to detect TD cases using information from patient medical records. Cases were identified by the presence of a TD ICD9/10 diagnosis code or a mention of specific tic keywords in their medical records (Supplemental Fig. 1). This approach identified 485 algorithm-defined TD cases, 316 of these were in the target sample. After a clinician chart review of these 316 charts, 266 were confirmed as true TD cases (84.2%).

This subset of clinically validated TD cases in BioVU served as a positive control for the TD PheRS (Supplemental Tables 11 and 12). We found that 75.6% (201/266) of the TD clinician-defined cases had PheRS scores within the tenth decile of the TD PheRS, and 91.4% (243/266) of the cases had PheRS scores falling in the seventh PheRS decile or higher (Fig. 4B). Additionally, we found that clinically validated cases had a significantly higher TD PheRS compared to non-cases according to the Wilcoxon rank-sum test ( $p < 2.2 \times 10^{-16}$ ) and logistic regression analysis accounting for covariates ( $p = 4.787 \times 10^{-151}$ ;  $\beta = 1.68$ ;  $SE = 0.06$ ; Fig. 4C).

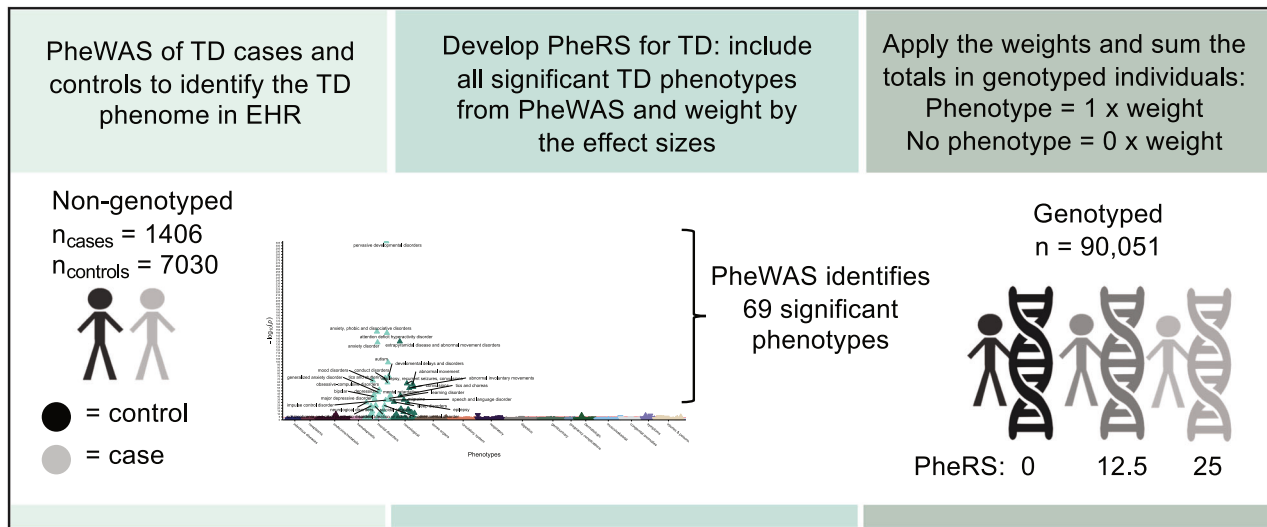
A small proportion of the clinically validated TD cases (23/266, 8.6%) had a very low TD PheRS, falling in the first three deciles (Fig. 4B). These individuals had no ICD9/10 codes in common with TD patients but had mentions of Tourette's, tics, chronic motor tic disorder, vocal and motor tics, or head tics in their medical records, indicating that these individuals likely visited the medical center for care independent of their tics but included a TD diagnosis within their medical histories. Because the TD PheRS relies solely on the presence or absence of ICD9/10 billing codes and not keywords in the EHR, these individuals had a much lower PheRS despite having a TD diagnosis. Conversely, there were multiple BioVU patients with a high TD PheRS that were not identified by the TD algorithm and clinician validation (825 individuals within the top percentile, Fig. 4B). These individuals share substantial overlapping phenotype with TD patients, despite few of them having a TD phecode (15.3%, 126/825). Lastly, within the top percentile of the TD PheRS, over 50% of patients had a broad range of diagnoses including malaise and fatigue, abdominal pain, anxiety disorder, convulsions, nausea and vomiting, depression, major depressive disorder, mood disorders, gastroesophageal reflux disease (GERD), and cough (Supplemental Table 13).

Because several neurodevelopmental disorders are highly comorbid with TD (i.e., ASD, ADHD, and OCD), we tested whether the TD PheRS was able to differentiate between the TD and ASD populations. We used a previously curated set of 444 ASD patients within BioVU. These individuals were identified by ICD9/10 codes and keywords within their charts and were verified by clinician chart review [53]. As expected, we found that the average TD PheRS was highest in the patients with both a TD and ASD diagnosis (mean  $\pm$  SD =  $2.57 \pm 0.76$ ) when compared to individuals with only a TD diagnosis (mean  $\pm$  SD =  $2.17 \pm 0.72$ ) or only an ASD diagnosis (mean  $\pm$  SD =  $1.85 \pm 0.67$ , Supplemental Fig. 2). We also found that the TD PheRS was significantly higher within the TD patients compared to the ASD patients ( $p = 2.25 \times 10^{-6}$ ;  $\beta = 0.32$ ;  $SE = 0.07$ ), suggesting that while there was substantial overlap, the TD PheRS was most sensitive to the clinically validated TD individuals (Supplemental Table 10).

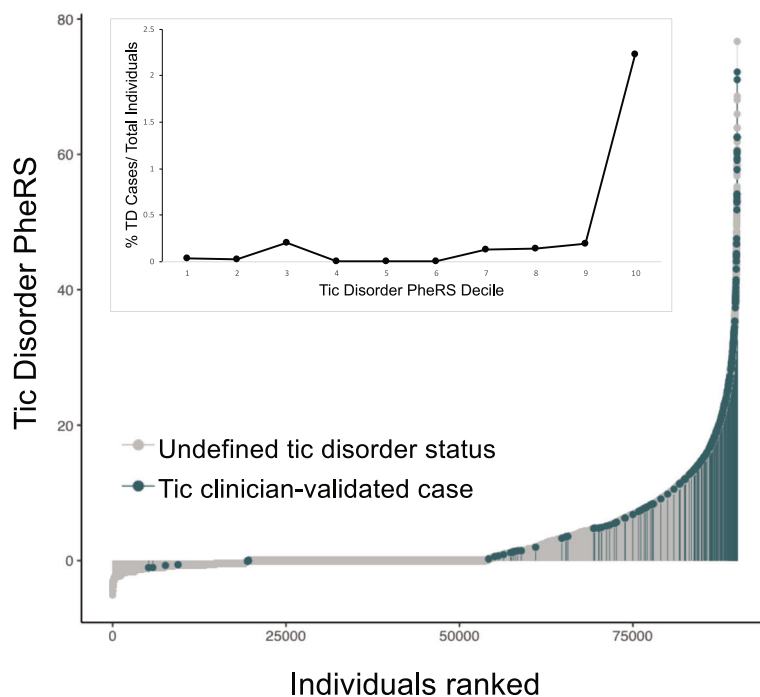
### DISCUSSION

The availability of de-identified medical records for research purposes provides an expedited and relatively low-cost way to study phenotypically complex disorders over the lifespan in large patient populations. In this study, we leveraged the large, clinical

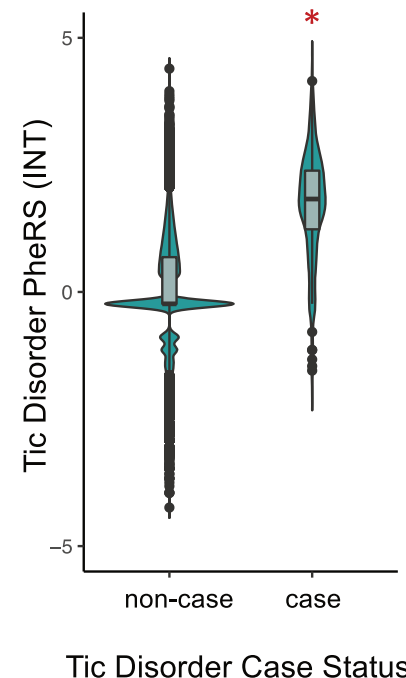
A.



B.



C.



**Fig. 4 TD Phenotype Risk Score (PheRS) identifies individuals with shared features of tic disorders.** **A** The TD PheRS was calculated for 90,051 individuals in the Vanderbilt biobank using the 69 phenotypes identified in the TD PheWAS. Each individual is given a 0 or 1 for the absence or presence of each phenotype and this value is multiplied by the weight of each phenotype (weight = effect estimate from PheWAS analysis). **B** The distribution of the TD PheRS across all genotyped individuals with clinician-validated cases highlighted in blue. All individuals not identified as cases by the TD algorithm and clinician review are gray. Inset shows the percentage of clinician-validated cases within each TD PheRS decile. **C** Violin plots with boxplot insets of the TD PheRS in clinician-validated cases and non-cases shows that the TD PheRS in cases is significantly higher compared to controls (Wilcoxon rank sum test,  $p < 2.2 \times 10^{-16}$ , logistic regression analysis  $p = 4.787 \times 10^{-151}$ ). The TD PheRS was inverse-normal transformed (INT) before plotting. In the logistic regression analysis, which was performed in the individuals of European ancestry, the covariates included PC1-10, current age, median age of medical record, number of medical center visits, and sex.

biobank and EHR database at Vanderbilt to evaluate the phenotypic complexity of TD. We identified TD cases using ICD9/10 diagnosis codes, matched controls, and tested 676 phenotypes within the EHR for enrichment or depletion among TD cases. Consistent with prior studies, we found complex overlapping phenotype between TD and several neuropsychiatric

phenotypes, with TD most often being the first psychiatric diagnosis in TD cases' medical records (~64%). Additionally, individuals diagnosed with TD in our sample received an average of 3.1 additional psychiatric diagnoses. Over half of the TD patients (51.9%) received at least two additional psychiatric diagnoses later in life. Comorbid psychiatric disorders often

negatively impact patients more than the tics themselves. Psychiatric comorbidities that occurred with increased prevalence in TD cases included anxiety disorders (29.6%), ADHD (19.1%), mood disorders (16.9%), depression (10.7%), ASD (9.4%), and OCD (8.7%) (Supplemental Table 3, Supplemental Table 5). These associations remained significant after conditioning on medication status (Supplemental Table 6). We also found that the associations of TD with lesser-known comorbidities, such as schizophrenia, personality disorders, and transient alteration of awareness, remained after adjusting for medication effects, a finding that warrants future study [54–57]. These findings emphasize the complex phenotypic trajectories for tic patients and support the need for thorough observation and follow-up by clinicians and caregivers even if tic symptoms subside.

Phenotype risk scores can be used in research to quantify the phenomic patterns of comorbidity that characterize TD. This single quantitative score can then be used to investigate broader phenotypic, and even genetic, liability to TD. Indeed, when phenotype prevalence is low, most people with high polygenic risk (i.e., a large number of alleles that are associated with tics) still will not have a TD. However, high genetic liability to TD also increases the odds of related diagnoses (ADHD, OCD, etc.). Thus, a score which quantifies the TD-phenome may provide a path forward to increasing sample sizes for genome-wide association studies of TD liabilities.

In this proof of principle study, we demonstrated that the TD PheRS is significantly higher in confirmed TD cases. Indeed, three quarters of the clinically validated TD cases had a PheRS in the highest decile. Of the 23 confirmed TD cases with a PheRS equal to zero or less, 6 did not meet the medical home criteria despite having mentions of a TD diagnosis in their medical records, making these cases difficult to identify without the use of natural language processing algorithms. For the remaining 17 individuals, despite frequent visits to the medical center and documentation of TD in the clinical notes, the TD phenotype was never coded. Additionally, there were several BioVU individuals in the top percentile of the TD PheRS lacking any TD phecode within their medical record (699/825, 84.7%). These may represent TD patients who are missing a diagnosis code for TD or may represent individuals with several features of TD, such as the several neuropsychiatric comorbidities, without the presence of tics. These findings represent the strength of the PheRS approach at identifying individuals who would have otherwise been labeled as controls and may confound a case-control study because they share many common features with TD patients. The strength of the PheRS lies in the ability to identify individuals with many features of a disease but without a formal diagnosis annotated in the EHR. It is possible that patients with a very high TD PheRS may reach this threshold with co-occurring diagnoses (i.e., ADHD, ASD, OCD) without ever experiencing tics; however, the PheRS approach allows us to identify these patients and review their medical history to develop a more accurate diagnosis. Additional research is needed to determine how informative PheRS phenotypes will be for GWAS of complex phenotypes like TD.

Despite the many conveniences of exploring phenotype information within the EHR, there are substantial limitations. The PheWAS is inherently biased by the number of medical center visits and number of ICD9/10 billing codes collected during these visits. High levels of missing data can exist in a patient's medical record. A missing diagnosis code does not necessarily reflect the absence of a phenotype and could instead be the result of a patient not receiving all of their care at a single hospital system, or miscoding by the medical center. For these reasons, running case-control studies based on a single diagnosis code can be skewed by misclassification of case and control status. In our approach we have attempted to correct for some of these biases by limiting the controls of our TD PheWAS to the "medical home" population, requiring five diagnosis codes on separate days over three

consecutive years, to enrich for individuals that receive the majority of their clinical care at Vanderbilt. Relying on ICD9/10 codes for diagnosis can be risky as institutional and provider coding biases within the EHR exist, especially among phenotypes that have been historically stigmatized, including mental health diagnoses, underscoring the importance of replicating these results in additional EHR systems and validating these approaches with clinician review. The PheRS approach can alleviate some of the EHR coding biases because the score is calculated across several phenotypes to identify patients that share multiple, distinct phenotypic features with the designated case population. Our analyses did not distinguish which Vanderbilt department or clinic individuals from the discovery or target populations visited, which could impact the number and types of codes they received (i.e., specialty neurology clinics versus general healthcare providers). It is also likely that the individuals diagnosed with TD at Vanderbilt represent more severely affected cases due to ascertainment bias of a tertiary care center and this is represented in the proportion of TD cases that received a psychiatric or neurological disorder as their first diagnosis within our medical system (47% of TD cases compared to 7% of controls, Supplemental Table 4). This finding underscores the importance of considering the types of codes being assigned to the case and control populations, as our control samples are enriched for childhood ailments (i.e. fractures, fevers, respiratory symptoms), which thus affects the phenotypes uncovered in the PheWAS which are used in the downstream PheRS construction. Generation of multiple PheRS's across diverse biobanks should be performed to assess the portability of these methods and examine the effects of different case and control populations.

Ultimately, we find that leveraging the dense phenotype information in a large, clinical biobank can replicate the phenotypic findings of prior studies of TD patients, while adding longitudinal medical outcomes. In addition to the utility of the biobank for research purposes, these tools may also benefit clinicians and TD patients. Our PheWAS reveals multiple comorbid psychiatric phenotypes with TD, which could help clinicians more thoroughly evaluate TD patients and provide necessary interventions and treatments in a timelier manner. The PheRS serves as a quantitative measure encapsulating the broad medical phenome for a given disease. Because the PheRS is diagnosis agnostic and can be applied to target populations with any sample size, this tool could be advantageous for evaluating disease risk based purely on the presentation of phenotypes and provides a phenome-wide perspective when evaluating complex disorders.

## METHODS

### The synthetic derivative is a database of de-identified electronic health records

The SD currently houses clinical information and documentation for over 3.6 million individuals who receive clinical care at Vanderbilt University Medical Center (VUMC) dating back to 1994. This information includes insurance billing codes (International Classification of Diseases, 9<sup>th</sup> and 10<sup>th</sup> editions/ICD-9 and ICD-10 codes, respectively), clinical procedure codes (Current Procedural Terminology/CPT codes), clinician notes, family histories, lab values, and prescribed medications [58].

### BioVU is a biorepository of genotype data linked to medical records

The Vanderbilt Institute for Clinical and Translational Research at VUMC curates BioVU, a clinical biorepository linked to the de-identified EHR information within the SD [59]. Patients seen at a Vanderbilt clinic are given the option to participate in the BioVU research program, which collects the leftover blood samples from routine clinical testing for genotyping and research. Sample collection for BioVU began in 2007 and is ongoing at Vanderbilt clinics across middle Tennessee. Currently, the BioVU biobank houses DNA samples linked to de-identified EHRs for 329,000 individuals.

### Identification of TD cases and controls in the EHR

We identified TD cases and controls within the EHR using the following criteria. TD cases were required to have at least two separate, temporally distinct instances of case inclusion TD phenotypes, defined by ICD-9 and ICD-10 (International Classification of Diseases, Ninth/Tenth Revision) billing codes (Supplemental Table 1). TD case inclusion codes are: tics (307.2), tic disorder (F95), tic disorder not otherwise specified (307.20), other tic disorders (F95.8), tic disorder unspecified (F95.9), transient tic disorder (307.21/F95.0), tics of organic origin (333.3), other tics of organic origin (G25.69), Tourette's disorder (307.23/F95.2), and chronic motor or vocal tic disorder (307.22/F95.1). Individuals were excluded from the analysis if any instance of TD exclusion codes (such as non-tic movement disorders) were present in the medical record (Supplemental Table 1). TD cases were restricted to the non-BioVU population within the synthetic derivative (SD) (to allow the BioVU sample to serve as the hold-out validation set), thus yielding 1,406 individuals in the SD who met TD case criteria. Controls were selected from the medical home population within the SD, defined as individuals who had visited a Vanderbilt clinic at least five times within a consecutive three-year period. Controls were restricted to the non-BioVU population within the SD and were age- and sex-matched to the TD cases (5 controls were matched to each case using current age and sex variables) with the MatchIt package in R [60]. Current age reflects the age of the patients at the time of data extraction for this study. Controls were excluded if any instances of either the TD inclusion or exclusion codes were present in the medical record, resulting in 7,030 TD controls. Independent-sample t-tests were performed to test for significant differences in demographic values between TD cases and controls. Chi-squared tests were performed to evaluate differences in proportions of EHR-reported ethnicity and race between cases and controls (Table 1).

### Phenome-wide association study for TD status

A phenome-wide association study (PheWAS) was performed to identify the medical phenotypes enriched in TD cases. TD cases ( $n = 1,406$ ) and matched controls ( $n = 7,030$ ) were assigned a one or a zero case status (exposure variable), and logistic regressions were performed across 676 phenotypes using the PheWAS package in R (versions 0.99.5–2 and 3.6.0, respectively) [61, 62]. The EHR phenotypes (outcome variables), were defined by phecodes mapped from ICD9/10 billing codes across all medical diagnoses as previously described [40]. Individuals were required to have at least two instances of an ICD9/10 code within the medical record to be considered a case for each outcome variable. A minimum of 20 cases was required to test each outcome. Logistic regression model covariates included sex, current age, median age within the medical record, EHR-reported race, EHR-reported ethnicity, and number of visits to the medical center. Phenotypes with an association  $p$ -value below the Bonferroni-corrected threshold ( $p = 7.396 \times 10^{-5}$ ,  $0.05/676$ ) were reported as significant. In a sensitivity analysis, we conditioned the TD PheWAS on the presence or absence of prescribed medications for tic disorders or common comorbidities (e.g., ASD or ADHD) from the medical records. The medications included in this sensitivity analysis were: amphetamine salts, aripiprazole, atomoxetine, botulinum toxin, bupropion, citalopram, clomipramine, clonidine, clonazepam, desipramine, dextmethylphenidate, dextroamphetamine, duloxetine, escitalopram, fluphenazine, fluoxetine, fluvoxamine, guanfacine, haloperidol, imipramine, lisdexamfetamine, methylphenidate, nortriptyline, olanzapine, paliperidone, paroxetine, pimozide, quetiapine, risperidone, sertraline, tetrabenazine, topiramate, venlafaxine, and ziprasidone [63–65].

### PheWAS phenotype group enrichment

The phecodes analyzed in the TD PheWAS were mapped from ICD9/10 billing codes and were grouped into 17 independent phenotype groups (e.g., psychiatric disorders, neurological conditions, and diseases of the circulatory system). The hypergeometric test was performed to identify phenotype groups that were over- or under-represented in the 69 significant PheWAS results. Phenotype groups with a  $p$ -value below the Bonferroni-corrected threshold ( $p = 0.00294$ ,  $0.05/17$ ) were considered significantly over- or under-represented. The fold enrichment or fold depletion was reported for each phenotype group in Fig. 3B.

### TD phenotype risk score construction

As described above, we performed a PheWAS for TD diagnosis and identified 69 phenotypes significantly enriched or depleted in the TD cases. These 69 phenotypes were then used to construct the TD PheRS in

the independent, genotyped BioVU population, or target sample. The following equation was used as previously described:

$$\text{PheRS}_i = \sum_{p=1}^m w_p x_{i,p}$$

where  $x_{i,p}$  is equal to 1 if individual  $i$  has phenotype  $p$  or 0 if the phenotype is not present in the individual's health records and  $w_p$  refers to the weight of phenotype  $p$ . The weights for each phenotype are equal to the effect size estimate of each phenotype from the initial PheWAS of the discovery sample. Therefore, phenotypes with the greatest association to TD status contribute most to the PheRS, whereas phenotypes with smaller associations contribute less. Supplemental Table 9 lists the 69 contributing phenotypes with their respective weights. The PheRS was then calculated in 90,051 BioVU individuals (Supplemental Table 8).

### Algorithm and clinician validation of TD cases in BioVU

A combination of TD diagnosis codes and keywords were used to define a TD algorithm that was applied to the BioVU sample. Inclusion criteria required at least two instances of TD ICD-9 billing codes within the EHR (Supplemental Fig. 1) or the single presence of the keywords *motor tic*, *vocal tic*, *Tourette*, or *tic disorder* in the clinical notes. Exclusion criteria included ICD-9 billing codes for muscular diseases (Supplemental Fig. 1). The TD algorithm identified 485 cases, 316 of which subsequently underwent clinician chart review because they overlapped with the target BioVU sample. Of the 316 algorithm cases, 266 were clinically validated as true TD cases (84.2%). The patients identified by the TD algorithm that was not confirmed by clinician chart review were excluded as a tic disorder case based on multiple criteria (e.g., having a broad mention of tics in the medical history without additional information, tics that appeared following medication use, and mentions of tic misdiagnosis after appearance of abnormal movements or seizures). Of the 266 clinically validated TD individuals, 95 met the ICD9/10 code criteria for TD case definition, 40 met the keyword criteria, while 131 met both.

### DATA AVAILABILITY

Synthetic derivative (de-identified electronic health records) and BioVU data are available from Vanderbilt University Medical Center with institutional restrictions that govern the acquisition, use, and dissemination of data. Individuals interested in using this data in a non-profit, academic setting can contact the Vanderbilt Institute for Clinical and Translational Research (research.support.services@vumc.org) and request an application to the Integrated Data Access and Services Core.

### REFERENCES

1. American Psychiatric Association, American Psychiatric Association (eds.). Diagnostic And Statistical Manual Of Mental Disorders: DSM-5. 5th edn. Washington, D.C.: American Psychiatric Association; 2013.
2. Front Matter. Principles and Practice of Movement Disorders. Elsevier, 2011. pp i–iii.
3. Scharf JM, Miller LL, Gauvin CA, Alabiso J, Mathews CA, Ben-Shlomo Y. Population prevalence of Tourette syndrome: a systematic review and meta-analysis. *Mov Disord*. 2015;30:221–8.
4. Knight T, Steeves T, Day L, Lowerison M, Jette N, Pringsheim T. Prevalence of Tic disorders: a systematic review and meta-analysis. *Pediatr Neurol*. 2012;47:77–90.
5. Cavanna AE, David K, Bandera V, Termine C, Balottin U, Schrag A, et al. Health-Related quality of life in gilles de la Tourette syndrome: a decade of research. *Behav Neurol*. 2013;27:83–93.
6. Evans J, Seri S, Cavanna AE. The effects of Gilles de la Tourette syndrome and other chronic tic disorders on quality of life across the lifespan: a systematic review. *Eur Child Adolesc Psychiatry*. 2016;25:939–48.
7. Kim S, Greene DJ, Bihun EC, Koller JM, Hampton JM, Acevedo H, et al. Provisional Tic Disorder is not so transient. *Sci Rep*. 2019;9:3951.
8. Comings DE, Comings BG. A controlled study of Tourette syndrome. I. Attention-deficit disorder, learning disorders, and school problems. *Am J Hum Genet*. 1987;41:701–41.
9. Spencer T, Biederman J, Harding M, O'Donnell D, Wilens T, Faraone S, et al. Disentangling the overlap between Tourette's disorder and ADHD. *J Child Psychol Psychiatry*. 1998;39:1037–44.
10. Stewart SE, Illmann C, Geller DA, Leckman JF, King R, Pauls DL. A controlled family study of attention-deficit/hyperactivity disorder and Tourette's disorder. *J Am Acad Child Adolesc Psychiatry*. 2006;45:1354–62.



11. Ruscio AM, Stein DJ, Chiu WT, Kessler RC. The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. *Mol Psychiatry*. 2010;15:53–63.
12. Mathews CA, Grados MA. Familiarity of Tourette syndrome, obsessive-compulsive disorder, and attention-deficit/hyperactivity disorder: heritability analysis in a large sib-pair sample. *J Am Acad Child Adolesc Psychiatry*. 2011;50:46–54.
13. Hirschtritt ME, Lee PC, Pauls DL, Dion Y, Grados MA, Illmann C, et al. Lifetime prevalence, age of risk, and genetic relationships of comorbid psychiatric disorders in Tourette syndrome. *JAMA Psychiatry*. 2015;72:325.
14. Yu D, Mathews CA, Scharf JM, Neale BM, Davis LK, Gamazon ER, et al. Cross-disorder genome-wide analyses suggest a complex genetic relationship between Tourette's syndrome and OCD. *Am J Psychiatry*. 2015;172:82–93.
15. Huisman-van Dijk HM, Van De Schoot R, Rijkeboer MM, Mathews CA, Cath DC. The relationship between tics, OC, ADHD and autism symptoms: A cross-disorder symptom analysis in Gilles de la Tourette syndrome patients and their family members. *Psychiatry Res*. 2016;237:138–46.
16. Browne HA, Hansen SN, Buxbaum JD, Gair SL, Nissen JB, Nikolajsen KH, et al. Familial clustering of Tic disorders and obsessive-compulsive disorder. *JAMA Psychiatry*. 2015;72:359.
17. Darrow SM, Grados M, Sandor P, Hirschtritt ME, Illmann C, Osiecki L, et al. Autism spectrum symptoms in a Tourette's disorder sample. *J Am Acad Child Adolesc Psychiatry*. 2017;56:610–7.e1.
18. Pauls DL, Fernandez TV, Mathews CA, State MW, Scharf JM. The inheritance of Tourette disorder: a review. *J Obsessive-Compuls Relat Disord*. 2014;3:380–5.
19. Davis LK, Yu D, Keenan CL, Gamazon ER, Konkashbaev AI, Derks EM, et al. Partitioning the heritability of tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS Genet*. 2013;9:e1003864.
20. Mataix-Cols D, Isomura K, Pérez-Vigil A, Chang Z, Rück C, Larsson KJ, et al. Familial risks of Tourette syndrome and chronic tic disorders: a population-based cohort study. *JAMA Psychiatry*. 2015;72:787.
21. Verkerk AJMH, Mathews CA, Joosse M, Eussen BHH, Heutink P, Oostra BA. Cntnap2 is disrupted in a family with Gilles de la Tourette syndrome and obsessive compulsive disorder. *Genomics*. 2003;82:1–9.
22. Lawson-Yuen A, Saldivar J-S, Sommer S, Picker J. Familial deletion within NLGN4 associated with autism and Tourette syndrome. *Eur J Hum Genet*. 2008;16:614–8.
23. Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, et al. Sequence variants in *SLITRK1* are associated with Tourette's syndrome. *Science*. 2005;310:317–20.
24. Ercan-Sencicek AG, Stillman AA, Ghosh AK, Bilguvar K, O'Roak BJ, Mason CE, et al. L-Histidine decarboxylase and Tourette's syndrome. *N Engl J Med*. 2010;362:1901–8.
25. Petek E, Windpassinger C, Vincent JB, Cheung J, Boright AP, Scherer SW, et al. Disruption of a novel gene (*IMMP2L*) by a breakpoint in 7q31 associated with Tourette syndrome. *Am J Hum Genet*. 2001;68:848–58.
26. Huang AY, Yu D, Davis LK, Sul JH, Tsetsos F, Ramensky V, et al. Rare copy number variants in *NRXN1* and *CNTN6* increase risk for Tourette syndrome. *Neuron*. 2017;94:1101–1111.e7.
27. Sundaram SK, Huq AM, Wilson BJ, Chugani HT. Tourette syndrome is associated with recurrent exonic copy number variants. *Neurology*. 2010;74:1583–90.
28. Liu S, Tian M, He F, Li J, Xie H, Liu W, et al. Mutations in *ASH1L* confer susceptibility to Tourette syndrome. *Mol Psychiatry*. 2020;25:476–90.
29. Depienne C, Ciura S, Trouillard O, Bouteiller D, Leitão E, Nava C, et al. Association of rare genetic variants in opioid receptors with Tourette syndrome. *Tremor Hyperkinetic Mov* 2019;9: <https://doi.org/10.7916/tohm.v0.693>.
30. Scharf JM, Yu D, Mathews CA, Neale BM, Stewart SE, Fageress JA, et al. Genome-wide association study of Tourette's syndrome. *Mol Psychiatry*. 2013;18:721–8.
31. Liu S, Yu X, Xu Q, Cui J, Yi M, Zhang X, et al. Support of positive association in family-based genetic analysis between *COL27A1* and Tourette syndrome. *Sci Rep*. 2015;5:12687.
32. Yu D, Sul JH, Tsetsos F, Nawaz MS, Huang AY, Yelaya I, et al. Interrogating the genetic determinants of Tourette's syndrome and other tic disorders through Genome-Wide Association Studies. *Am J Psychiatry*. 2019;176:217–27.
33. Georgitsi M, Willsey AJ, Mathews CA, State M, Scharf JM, Paschou P. The genetic etiology of Tourette syndrome: large-scale collaborative efforts on the precipice of discovery. *Front Neurosci*. 2016;10:351.
34. Denny JC. Chapter 13: mining electronic health records in the genomics era. *PLoS Comput Biol*. 2012;8:e1002823.
35. Bastarache L, Hughey JJ, Hebbiring S, Marlo J, Zhao W, Ho WT, et al. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science*. 2018;359:1233–9.
36. Bastarache L, Hughey JJ, Goldstein JA, Bastraache JA, Das S, Zaki NC, et al. Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. *J Am Med Inform Assoc*. 2019;26:1437–47.
37. Lebovitch DS, Johnson JS, Dueñas HR, Huckins LM. Phenotype Risk Scores: moving beyond 'cases' and 'controls' to classify psychiatric disease in hospital-based biobanks. *Psychiatry Clin Psychol*. 2021. <https://doi.org/10.1101/2021.05.25.21249615>.
38. Wendt FR, Pathak GA, Deak JD, De Angelis F, Koller D, Cabrera-Mendoza B, et al. Using phenotype risk scores to enhance gene discovery for generalized anxiety disorder and posttraumatic stress disorder. *Mol Psychiatry*. 2022;27:2206–15.
39. Shaw DM, Polikowsky HP, Pruett DG, Chen H-H, Petty LE, Viljoen KZ, et al. Phenome risk classification enables phenotypic imputation and gene discovery in developmental stuttering. *Am J Hum Genet*. 2021;108:2271–83.
40. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-11-CM codes to phecodes: workflow development and initial evaluation. *JMIR Med Inform*. 2019;7:e14325.
41. Comings BG, Comings DE. A controlled study of Tourette syndrome. V. Depression and mania. *Am J Hum Genet*. 1987;41:804–21.
42. Kwak C, Vuong KD, Jankovic J. Migraine headache in patients with Tourette syndrome. *Arch Neurol*. 2003;60:1595.
43. Kirov R, Becker A, Rothenberger A. Sleep in Tourette syndrome. *Curr Dev Disord Rep*. 2014;1:252–9.
44. Hibberd C, Charman T, Bhatia RS, Tekes S, Hedderly T, Gringras P, et al. Sleep difficulties in children with Tourette syndrome and chronic tic disorders: a systematic review of characteristics and associated factors. *Sleep*. 2020;43:zs308.
45. Kompolti K, Goetz CG. Hyperkinetic movement disorders misdiagnosed as tics in Gilles de la Tourette syndrome. *Mov Disord*. 1998;13:477–80.
46. Damásio Joana, Edwards MarkJ, Alonso-Canovas Araceli, Schwingsenschuh Petra, Kägi Georg, Bhatia KailashP. The clinical syndrome of primary tic disorder associated with dystonia: a large clinical series and a review of the literature. *Mov Disord*. 2011;26:679–84.
47. Stone L, Jankovic J. The coexistence of tics and dystonia. *JAMA Neurol*. 1991;48:862–5.
48. Burd L, Kerbeshian J, Barth A, Klug MG, Avery K, Benz B. Long-term follow-up of an epidemiologically defined cohort of patients with Tourette syndrome. *J Child Neurol*. 2001;16:431–7.
49. Schlander M, Schwarz O, Rothenberger A, Roessner V. Tic disorders: administrative prevalence and co-occurrence with attention-deficit/hyperactivity disorder in a German community sample. *Eur Psychiatry*. 2011;26:370–4.
50. Lewin AB, Murphy TK, Storch EA, Conelea CA, Woods DW, Scahill LD, et al. A phenomenological investigation of women with Tourette or other chronic tic disorders. *Compr Psychiatry*. 2012;53:525–34.
51. Lichter DG, Finnegan SG. Influence of gender on Tourette syndrome beyond adolescence. *Eur Psychiatry*. 2015;30:334–40.
52. Girgis J, Martino D, Pringsheim T. Influence of sex on tic severity and psychiatric comorbidity profile in patients with pediatric tic disorder. *Dev Med Child Neurol*. 2022;64:488–94.
53. Singer EV, Niarchou M, Maxwell-Horn A, Hucks D, Johnston R, Sutcliffe JS, et al. Characterizing sleep disorders in an autism-specific collection of electronic health records. *Sleep Med*. 2022;92:88–95.
54. Müller N, Riedel M, Zawta P, Günther W, Straube A. Comorbidity of Tourette's syndrome and schizophrenia—biological and physiological parallels. *Prog Neuropsychopharmacol Biol Psychiatry*. 2002;26:1245–52.
55. Kerbeshian J, Peng C-Z, Burd L. Tourette syndrome and comorbid early-onset schizophrenia. *J Psychosom Res*. 2009;67:515–23.
56. Robertson MM, Banerjee S, Fox-Hiley PJ, Tannock C. Personality disorder and psychopathology in Tourette's syndrome: A controlled study. *Br J Psychiatry*. 1997;171:283–6.
57. Cavanna AE, Robertson MM, Critchley HD. Schizotypal personality traits in Gilles de la Tourette syndrome. *Acta Neurol Scand*. 2007;116:385–91.
58. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary use of clinical data: the Vanderbilt approach. *J Biomed Inform*. 2014;52:28–35.
59. Roden D, Pulley J, Basford M, Bernard G, Clayton E, Balser J, et al. Development of a large-scale de-identified DNA Biobank to enable personalized medicine. *Clin Pharmacol Ther*. 2008;84:362–9.
60. Ho D, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Softw*. 2011;42:1–28.
61. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 2010;26:1205–10.
62. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*. 2014;30:2375–6.
63. Walkup, JT. A Guide to Tourette Syndrome Medications. Tourette Syndrome Association, Inc.; 2008.
64. Jankovic J. Botulinum toxin in the treatment of dystonic tics. *Mov Disord*. 1994;9:347–9.
65. Kwak CH, Hanna PA, Jankovic J. Botulinum toxin in the treatment of tics. *Arch Neurol*. 2000;57:1190–3.

## ACKNOWLEDGEMENTS

The authors acknowledge helpful discussions during the development of the tic disorder phenotype risk score and EHR algorithm, provided by the Psychiatric Genetic Consortium Tourette Syndrome Workgroup and the Tourette Syndrome Workgroup at Vanderbilt University Medical Center. We thank Maria Niarchou for revising the manuscript and providing feedback on the analyses. The Synthetic Derivative and BioVU projects at VUMC are supported by numerous sources, including the NIH-funded Shared Instrumentation Grant S10OD017985 and S10RR025141; CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975 from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent official views of the National Center for Advancing Translational Sciences or the National Institutes of Health. Genomic data are also supported by investigator-led projects that include U01HG004798, R01NS032830, RC2GM092618, P50GM115305, U01HG006378, U19HL065962, R01HD074711; and additional funding sources listed at <https://vict.vumc.org/biovu-funding/>. TWM receives funding from the Tourette Association of America and is supported by the National Human Genome Research Institute (5T32HG008341). JMS receives grant support from the TLC Foundation for Body-Focused Repetitive Behaviors. JMS, CAM, and LKD are supported by grants from the National Institute of Neurological Disorders and Stroke, NS105746 and NS102371. DAI receives funding from the Tourette Association of America, Vanderbilt Clinical and Translational Research Scholars (KL2TR002245), National Institute of Neurological Disorders and Stroke (K23NS131592).

## AUTHOR CONTRIBUTIONS

TWM and LKD conceptualized and designed the study. TWM, AA, and EG performed the analyses. DY, DAI, CAM, and JMS provided clinical, intellectual, and methodological insights. All authors contributed to interpreting results, writing, reviewing, and editing the manuscript. All authors approved the final manuscript.

## COMPETING INTERESTS

JMS is an unpaid member of the Scientific Advisory Board of the Tourette Association of America. DAI receives funding from Teva Branded Pharmaceutical Products, R&D, Inc. The other authors declare no competing interests.

## ETHICAL APPROVAL

This study complies with all relevant ethical regulations as approved by the VUMC institutional review board (IRB) and designated as non-human participants research because of the use of fully de-identified data (IRB 190418).

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41398-024-03011-w>.

**Correspondence** and requests for materials should be addressed to Tyne W. Miller-Fleming or Lea K. Davis.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024