nature genetics

ANALYSIS

https://doi.org/10.1038/s41588-018-0313-7

Corrected: Author Correction

Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes

Chirag M. Lakhani[1], Braden T. Tierney[1,2], Arjun K. Manrai[1,3], Jian Yang[4,5], Peter M. Visscher[4,5,6*] and Chirag J. Patel[1,6*]

We analysed a large health insurance dataset to assess the genetic and environmental contributions of 560 disease-related phenotypes in 56,396 twin pairs and 724,513 sibling pairs out of 44,859,462 individuals that live in the United States. We estimated the contribution of environmental risk factors (socioeconomic status (SES), air pollution and climate) in each phenotype. Mean heritability ($h^2 = 0.311$) and shared environmental variance ($c^2 = 0.088$) were higher than variance attributed to specific environmental factors such as zip-code-level SES ($var_{SES} = 0.002$), daily air quality ($var_{AQI} = 0.0004$), and average temperature ($var_{temp} = 0.001$) overall, as well as for individual phenotypes. We found significant heritability and shared environment for a number of comorbidities ($h^2 = 0.433$, $c^2 = 0.241$) and average monthly cost ($h^2 = 0.290$, $c^2 = 0.302$). All results are available using our Claims Analysis of Twin Correlation and Heritability (CaTCH) web application.

Disentangling how genetic and environmental factors contribute to many phenotypes in the same population has been largely unfeasible to date. Most study designs consider a single disease or environmental factor at a time. Administrative health data, such as insurance claims and electronic health records, may enable more comprehensive analyses of the roles of genetics and shared environment in hundreds of phenotypes. Here, we analysed a massive, individual-level claims dataset of 44,859,462 individuals to systematically partition phenotypic variance between genetic and non-genetic factors across a large US population. Documenting both the genetic and environmental contributions of phenotypic variance is instrumental for major health studies, such as the United States' All of Us effort[1,2]. Furthermore, the use of genome sequence data in medical decisionmaking is under debate[2] and estimating heritability in a 'real-world' setting can help to quantify the clinical utility of genome sequencing[3].

In human genetics, heritability is defined as the amount of phenotype or disease variation that can be attributed to genetic factors. In family studies, other important quantities, such as 'shared environment' and 'non-shared environment', complement heritability and describe variation in phenotype resulting from non-genetic factors. Estimation of heritability and environmental components of phenotypic variation have historically used family-based studies, such as those involving twins that are concordant (and discordant) for disease. However, building twin registries can be resource-intensive in the ascertainment of both twin pairs and phenotypes. What is missing are family-based studies that measure numerous phenotypes across a large and diverse population that experience a variety of environmental exposures. First, health administration data enable such an approach because these data give a comprehensive snapshot of health (for example, thousands of disease diagnoses and laboratory reports, in addition to the cost of healthcare), and they enable family-based[4,5] or twin-based studies across a large number of diseases. Although twin-based analysis in such datasets is difficult because of a lack of zygosity information, we employed methodology[6] that utilizes sex information to differentiate between identical and non-identical twin pairs.

Second, there has also been a great deal of interest in understanding the contribution of one's residence or 'zip code' in their disease state[7,8]. Individual-level data with geographical and temporal information (that is, patient mailing zip code and time of diagnosis) can enable an understanding of the contribution of specific geographically linked environmental factors in phenotypic variation. To our knowledge, only one study has attempted to quantify the relative contribution of local environment and genetics[9]. In our analysis, we quantify the relative contribution of local environment and genetics by integrating individual-level data with zip code-level information that serve as geographical indicators of the area's SES, air pollution quality level and weather/climate.

We estimated heritability and shared environmental variance for 560 phenotypes (based on diagnostic billing codes and laboratory tests) in a cohort of 56,396 twin pairs born on or after 1985 (individuals that are on their parent's/guardian's insurance plan) using an individual-level claims dataset of 44,859,462 individuals from the United States. We also estimated phenotypic correlation for same sex and opposite sex siblings using a cohort of 724,513 sibling pairs (Supplementary Note). We estimated the contribution of specific environmental risk factors, such as SES, air pollution, and climate difference, to these phenotypes by linking individual claimants to external datasets via residential locations (Fig. 1d–g). In addition, we computed genetic and environmental contributions to the cost of care utilization and total comorbidities. Finally, we estimated the validity of our estimates for heritability and shared environment through systematic comparison of documented estimates in the published literature.

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [2]Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA, USA. [3]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. [4]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia. [5]Queensland Brain Institute, The University of Queensland, Brisbane, Australia. [6]These authors jointly supervised this work: Peter M. Visscher, Chirag J. Patel. *e-mail: peter.visscher@uq.edu.au; chirag_patel@hms.harvard.edu

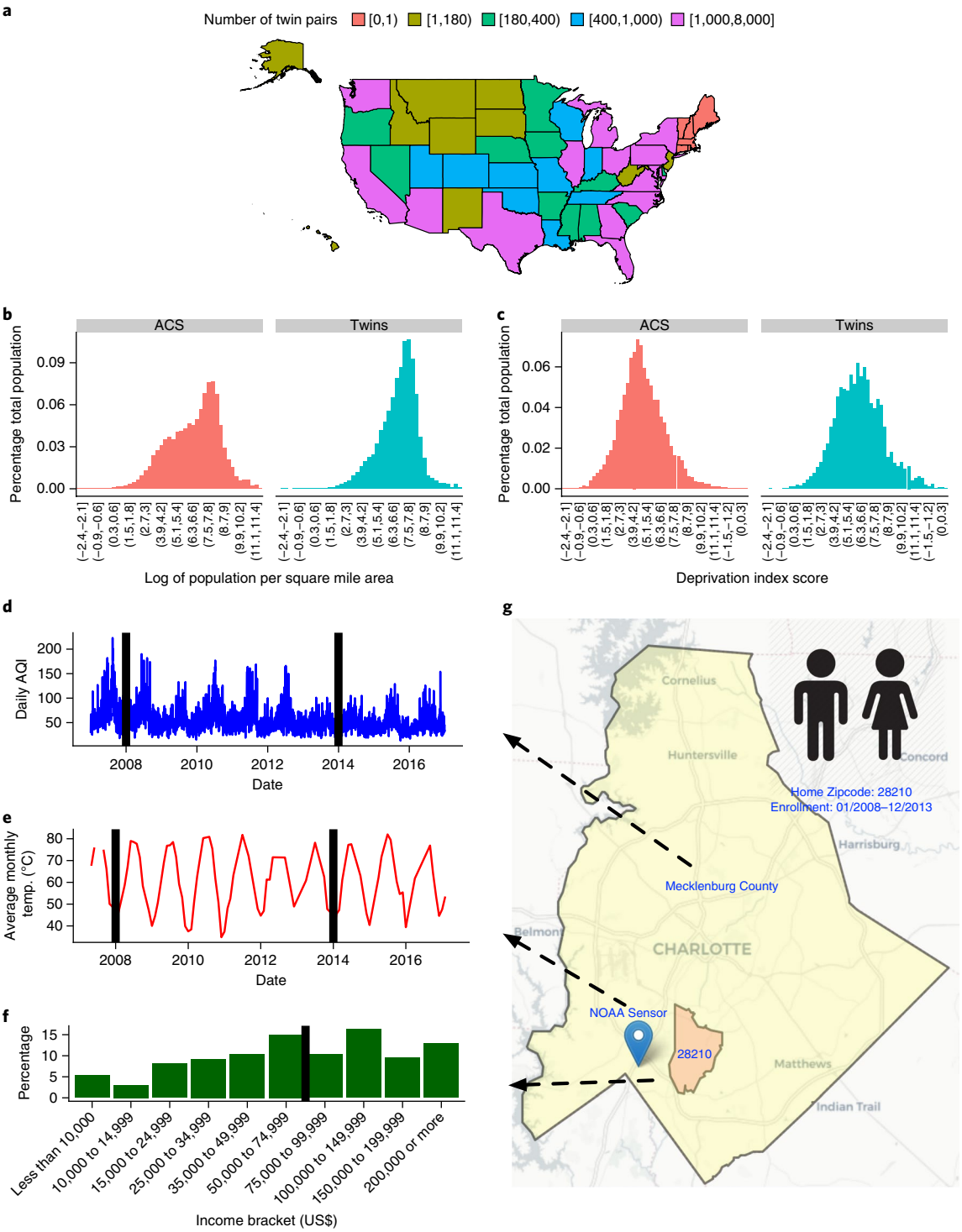

Fig. 1 | Geographic distribution of 56,396 twin pairs in CaTCH and an example of environmental data aggregation on a zip code basis. a, Count of twin pairs in CaTCH for each state in the United States. **b**, Distribution of log of population density for the entire United States (based on Census American Community Survey data) and twin pairs. **c**, Distribution of deprivation index for the entire United States and twin pairs. **d**, Time series for daily AQI for Mecklenburg county. Black lines represent the years 2008 and 2014. **e**, Time series for average monthly temperature for NOAA sensor closest to zip code 28210. Black lines represent the years 2008 and 2014. **f**, Distribution of median family income distribution among residents of zip code 28210. Black line represents the mean median income value. **g**, Map of county, zip code, and closest NOAA sensor for hypothetical twin pair residing in zip code 28210. Background map image from OpenStreetMap licensed under the terms of the Creative Commons Attribution-ShareAlike 2.0 license (CC BY-SA). ACS, American Community Survey; NOAA, National Oceanic and Atmospheric Administration.

Results

Data overview. We utilized de-identified member claims data from Aetna Inc., a national health insurance company, to assemble a cohort of 56,396 twin pairs and 724,513 sibling pairs (Methods and Supplementary Note) that were members for at least 3 years in the entire surveillance period between 01/01/2008 and 01/02/2016 .

Table 1 | Characteristic of ascertained insurance claims twin and sibling cohorts

	All pairs	FF pairs	MM pairs	MF pairs
Number of twin pairs	56,396	17,835	17,919	20,642
Number of sibling pairs	724,513	171,095	187,033	366,385
Median age at start of surveillance (IQR) (twin)	7 (3-13)	8 (3-14)	8 (3-13)	7 (2-12)
Median age at start of surveillance (IQR) (sibling)	7 (2-12)	7 (2-12)	7 (2-12)	7 (2-12)
Median months of surveillance (IQR) (twin)	60 (45-84)	60 (45-84)	60 (45-84)	60 (45-84)
Median months of surveillance (IQR) (sibling)	61 (46-84)	61 (46-84)	61 (46-84)	61 (46-84)
Median number of ICD Codes (IQR) (twin)	23 (12-42)	23 (12-41)	22 (11-41)	24 (13-44)
Median number of ICD Codes (IQR) (sibling)	23 (12-42)	24 (13-42)	22 (11-41)	23 (12-42)
Distinct number of zip codes (twin)	11,666	7,302	7,235	7,466
Distinct number of zip codes (sibling)	24,703	17,324	17,606	21,112
Surveillance period	01/01/2008 - 01/02/2016			

FF pairs, twin pairs where both individuals are female; MM pairs, twin pairs where both individuals are male; MF pairs, twin pairs where one individual is male and the other is female; IQR, interquartile range.

The median age of twin and sibling pairs at the start of surveillance was 7 years (Table 1). The age range for twins and siblings in this cohort was between 0 and 24 years. Using the claims data, we mapped health claims codes to higher level phenotypes called phenome-wide association studies (PheWAS) codes[10] (Methods). Phenotypic filtering produced 551 PheWAS codes, seven quantitative phenotypes, and two derived quantitative phenotypes. The twin cohort was geographically heterogeneous. There were 38 states with at least 100 twin pairs, whereas six states had no twin pairs (Fig. 1a). Overall, the twin pairs resided in areas with higher income and population density (Fig. 1b,c). The prevalence of PheWAS phenotypes among twin pairs was variable within and between different functional domains (prevalence = 0.30–73.2%) (Supplementary Fig. 1). All results, including phenotype specific data, are available using our CaTCH web application (see URLs).

Estimation of h^2 and c^2. We used a twin-based method to estimate the proportion of phenotypic variance resulting from additive genetic factors (that is, the narrow-sense heritability, h^2) and variance resulting from environmental factors shared between twins (c^2). Given the lack of zygosity information, we estimated h^2 and c^2 using the difference in correlation between same sex (r_{twinSS}) and opposite sex twin pairs (r_{twinOS}), assuming that opposite sex pairs are dizygotic and same sex twin pairs are a mixture of monozygotic and dizygotic twin pairs (Methods). We tested the validity of the assumption that r_{twinOS} is a good proxy for same sex dizygotic twin correlation (r_{twinDZSS}) by creating a non-twin sibling cohort and estimating the correlation between same sex sibling correlation (r_{sibSS}) and opposite sex sibling correlation (r_{sibOS}) for all 551 binary phenotypes (Supplementary Note). We found r_{sibSS} and r_{sibOS} were highly correlated ($r = 0.978$, 95% CI: 0.974, 0.981) (Supplementary Fig. 2). Also, for 95% of phenotypes, $r_{\text{sibSS}} - r_{\text{sibOS}}$ ranged between −0.012 and 0.051 and r_{sibSS} was, on average, 0.017 higher than r_{sibOS} (Supplementary Fig. 3), but for 23.5% of phenotypes $r_{\text{sibSS}} - r_{\text{sibOS}}$ followed the null distribution (pi_0 statistic[11]). We conclude that r_{twinOS} is highly correlated with r_{twinDZSS} for these 551 phenotypes. However, we found that r_{twinOS} is slightly lower, on average, than r_{twinDZSS}. Therefore, the estimates of h^2 and c^2 will be slightly biased. We also found r_{twinOS} is, in general, larger than both r_{sibOS} and r_{sibSS} (Supplementary Fig. 4). Therefore, using r_{sibSS} instead of r_{twinOS} as a proxy for r_{twinDZSS} replaces one biased estimator for another (Supplementary Note). We also found strong evidence to the validity of our assumption of Weinberg's Rule (Supplementary Note).

Overall phenome-wide summary of h^2 and c^2. The inverse-variance weighted mean estimate among all phenotypes was 0.316 (95% CI: 0.296, 0.335) for h^2 and 0.088 (95% CI: 0.074, 0.102) for c^2 (Fig. 2a). In addition, among all phenotypes, the opposite and same sex correlations for twins ($r_{\text{twinSS}} = 0.307$, 95% CI: 0.297, 0.318, $r_{\text{twinOS}} = 0.240$, 95% CI: 0.229, 0.251) were higher than for the siblings ($r_{\text{sibSS}} = 0.199$, 95% CI: 0.192, 0.206, $r_{\text{sibOS}} = 0.182$, 95% CI: 0.175, 0.189). The r_{twinSS} estimate was highest because same sex twin pairs are a mixture of monozygotic and dizygotic twin pairs. The higher value for r_{twinOS} compared to both r_{sibSS} and r_{sibOS} was a result of larger twin shared environment versus the sibling shared environmental effect.

Accounting for multiple hypotheses by controlling the false discovery rate (FDR) at 5%, we found 326/560 (58.2%) phenotypes had a non-zero heritability ($h^2 > 0$) and 180/560 (32.1%) phenotypes had non-zero shared environmental effects ($c^2 > 0$). Of these phenotypes, 225/560 (40%) h^2 estimates and 138/560 (24.6%) c^2 estimates remained significant at a more stringent significance level by Bonferroni-adjusted $P < 0.05$. We show a volcano plot of both h^2 and c^2 estimates for all 560 phenotypes, where the dotted line represents the FDR threshold for each statistic (Fig. 2b,c). The majority of age (β_{age}) and sex (β_{sex}) fixed effects were also non-zero (Methods and equation (2)). Controlling for multiple hypotheses using an FDR threshold of 0.05 there were 487/560 (86.9%) phenotypes for β_{age} and 281/560 (50.1%) phenotypes for β_{sex} that were FDR significant, respectively (see URLs).

Among functional domains with at least five phenotypes, the domains with the highest h^2 were quantitative laboratory measures ($h^2 = 0.799$, 95% CI: 0.551,1.048, seven out of seven phenotypes reached FDR threshold) and cognitive ($h^2 = 0.594$, 95% CI: 0.355, 0.834, four out of five phenotypes reached FDR threshold) (Fig. 2a). The lowest were connective tissue ($h^2 = 0.170$, 95% CI: 0.108, 0.233, 2 out of 11 phenotypes reached FDR threshold) and environment ($h^2 = 0.211$, 95% CI: 0.161, 0.260, 24 out of 45 phenotypes reached FDR threshold) (Fig. 2a).

The functional domains with the highest c^2 were ophthalmological ($c^2 = 0.183$, 95% CI: 0.147, 0.218, 27 out of 42 phenotypes reached FDR threshold) and respiratory ($c^2 = 0.182$, 95% CI: 0.151, 0.213, 34 out of 48 phenotypes reached FDR threshold) (Fig. 2a). The lowest were reproduction ($c^2 = -0.073$ 95% CI: −0.146, 0.000, three out of 18 phenotypes reached FDR threshold) and cognitive ($c^2 = -0.048$, 95% CI: −0.145, 0.049, two out of five phenotypes reached FDR threshold) (Fig. 2a)

From all 560 phenotypes in this study, there were 294 phenotypes (52.5%) in which c^2 followed the null distribution (pi_0 statistic[11]) (Methods), consistent with a model where twin resemblance was solely a result of additive genetic variance.

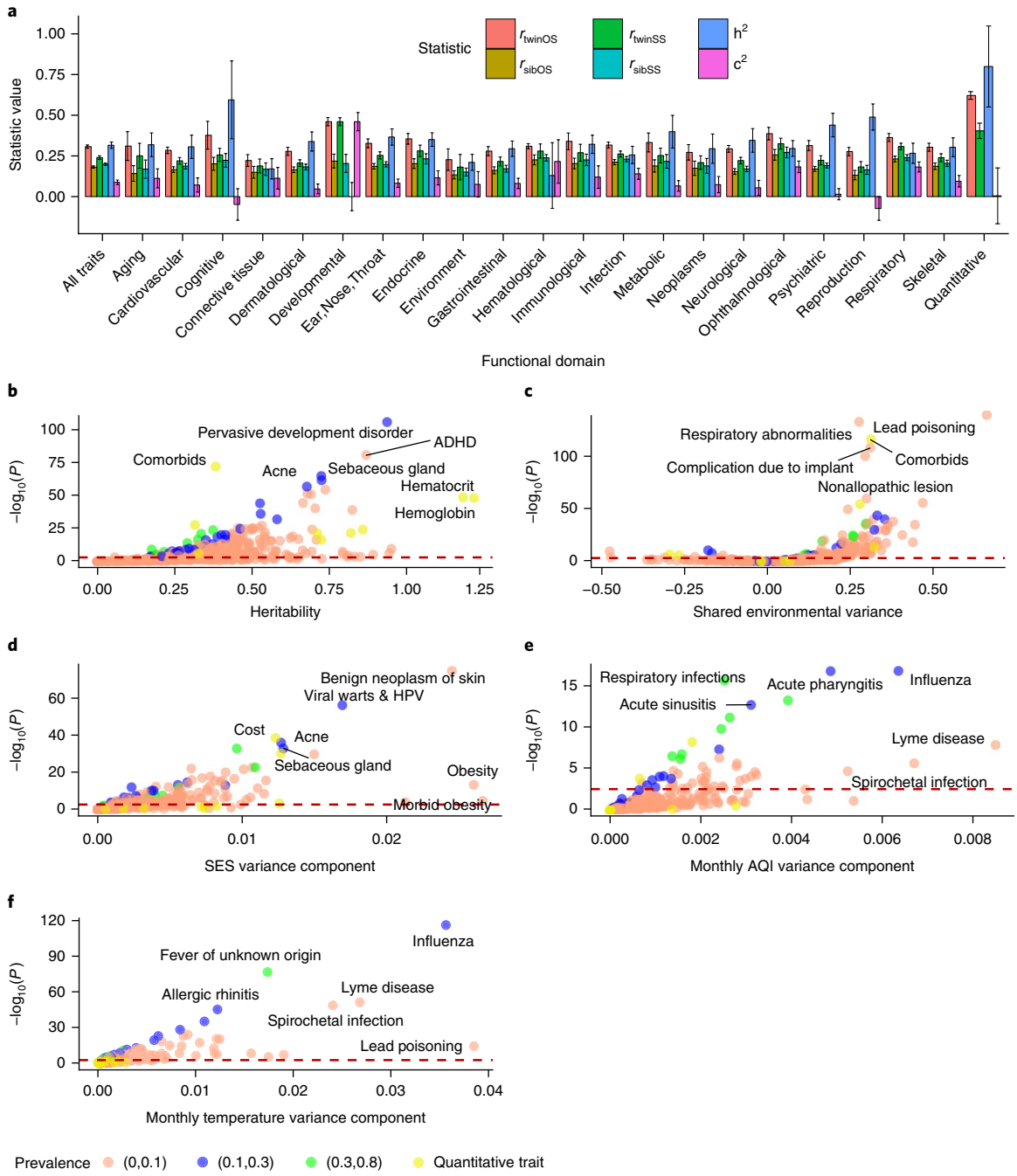

Fig. 2 | Estimates of twin statistics across functional domains and individual basis for 56,396 twin pairs in CaTCH among all 560 phenotypes. a, Barplot of meta-analytic estimates of r_{twinOS}, r_{twinSS}, h^2, and c^2 among all 560 phenotypes and within functional domains (error bars represent 95% CI). **b**–**f,** Volcano plots for estimates of h^2, c^2, var_{SES}, var_{AQI}, and var_{temp}, along with labels for phenotypes with top *P* values and large effect sizes for each estimate. Dashed red lines represent the threshold for Benjamini–Yekutieli FDR adjusted *P* values passing significance ($P = 0.05$) for each estimate.

Cost and comorbidities have significant h^2 and c^2. We found that average monthly cost had both significant $h^2 > 0$ and $c^2 > 0$ (Fig. 3b) in the twin pairs. Specifically, the estimate of h^2 was 0.290 (95% CI: 0.241, 0.339) and 0.433 (95% CI: 0.390, 0.477) for average monthly cost and number of PheWAS comorbidities, respectively. Estimates of c^2 were comparable; $c^2 = 0.302$, 95% CI: 0.271, 0.332 for average monthly cost and $c^2 = 0.241$, 95% CI: 0.213, 0.268 for number of PheWAS comorbidities (Fig. 3b). The same and opposite sex twin correlations (r_{twinSS} and r_{twinOS}) for number of PheWAS comorbidities ($r_{twinSS} = 0.549$, 95% CI: 0.543, 0.556, $r_{twinOS} = 0.458$, 95% CI: 0.450, 0.465) were slightly higher than average monthly claims cost ($r_{twinSS} = 0.508$, 95% CI: 0.501, 0.515, $r_{twinOS} = 0.447$, 95% CI: 0.439, 0.455) (Fig. 3b).

Specific geocoded environmental factors. In the same model, we estimated the proportion of variance in a phenotype attributable to environmental risk factors (based on home zip code), including an SES 'index' (Supplementary Note) (var_{SES}), median air quality index exposure (var_{AQI}), and median monthly average temperature exposure (var_{temp}) in addition to h^2 and c^2. The variance components for environmental risk factors were modest compared to h^2 and c^2. For all phenotypes, $var_{SES} = 0.002$ (95% CI: 0.002, 0.002), $var_{AQI} = 0.0001$

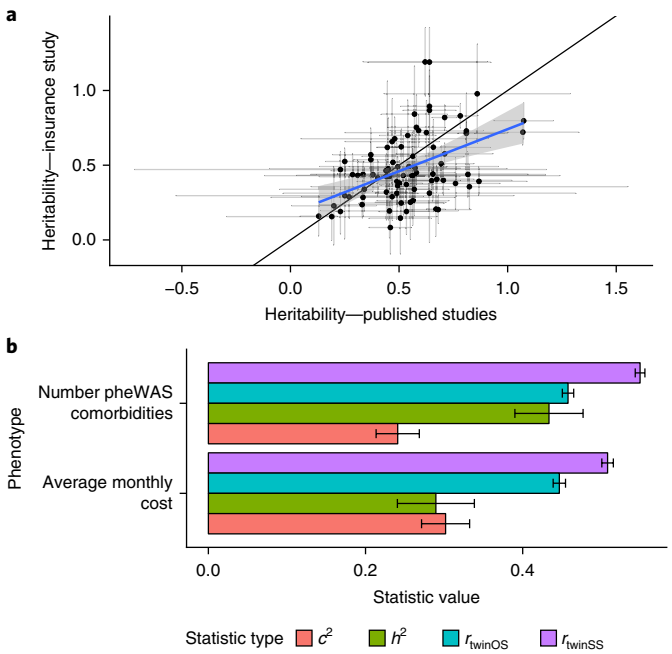

Fig. 3 | Comparison of h^2 estimates in CaTCH to published literature and estimates for cost and comorbidities in CaTCH. a, Scatterplot of published h^2 estimates from 56,396 twin pairs in CaTCH versus h^2 estimates from 81 published studies; vertical and horizontal error bars represent 95% CI for CaTCH and published estimates, respectively, black line is line with slope 1 and intercept 0, blue line is line of best fit and grey shaded region is 95% CI for line of best fit. **b**, Barplot of estimates of h^2, c^2, r_{twinOS}, and r_{twinSS} for the phenotypes average monthly cost and number of PheWAS comorbidities from 56,396 twin pairs; error bars represent 95% CI.

(95% CI: 0.0003, 0.0005), and var_{temp} = 0.001 (95% CI: 0.001, 0.001) were much smaller than the mean estimates of h^2 and c^2 described earlier (Supplementary Fig. 5). Controlling for multiple hypotheses using an FDR threshold of 0.05, we found 145/560 phenotypes for var_{SES}, 36/560 phenotypes for var_{AQI}, and 117/560 phenotypes for var_{temp} that passed FDR significance. Phenotypes with the largest var_{SES} were morbid obesity (var_{SES} = 0.027, 95% CI: 0.014, 0.039) and benign neoplasm of skin (var_{SES} = 0.024, 95% CI: 0.022, 0.027). Phenotypes with the largest var_{AQI} were Lyme disease (var_{AQI} = 0.008, 95% CI: 0.006, 0.011) and average monthly cost (var_{AQI} = 0.006, 95% CI: 0.004, 0.009). Phenotypes with the largest var_{temp} were lead poisoning (var_{temp} = 0.039, 95% CI: 0.029, 0.048) and influenza (var_{temp} = 0.036, 95% CI: 0.033, 0.039) (Fig. 2d–f).

Comparison to published literature. We compared our estimates of h^2 and c^2 to a large meta-analysis of twin studies[12] (meta-analysis of twin correlations and heritability, MaTCH) containing 9,568 phenotypes from 5,169,879 twin pairs where monozygotic and dizygotic correlations were reported. The two major differences between CaTCH and MaTCH were that CaTCH studied 38 infectious diseases compared with MaTCH and that the CaTCH cohort was younger than most of the studies in MaTCH (Supplementary Note).

Comparing the CaTCH estimates to MaTCH estimates, we observed that mean claims heritability (h^2 = 0.315, 95% CI: 0.296, 0.334) was smaller than the mean MaTCH estimate (h^2 = 0.593, 95% CI: 0.577, 0.608) (Fig. 4a). Furthermore, the mean CaTCH shared environment (c^2 = 0.088, 95% CI: 0.074, 0.102) was higher than the mean MaTCH estimate (c^2 = 0.042, 95% CI: 0.028, 0.055) (Fig. 4b)[12]. Comparing CaTCH h^2 estimates with MaTCH h^2 estimates along functional domains, we observed overlap between the 95% CI from h^2 CaTCH estimates and 95% CI from h^2 MaTCH estimates for 7 out of 21 functional domains, namely cognitive, endocrine, environment, hematological, infection, psychiatric, and reproduction functional domains (Fig. 4a). For c^2, the 95% CI from CaTCH estimates overlapped with the 95% CI from the MaTCH estimates for 11 out of 21 functional domains, namely cardiovascular, dermatological, endocrine, gastrointestinal, hematological, immunological, infection, metabolic, psychiatric, reproduction, and skeletal functional domains (Fig. 4b). In the MaTCH analysis, 69.1% of phenotypes were consistent with a model where twin resemblance was solely a result of additive genetic variance[12] compared with 52.5% of phenotypes in CaTCH.

Although we observed differences in heritability between CaTCH and MaTCH for aggregate phenotypic categories, we observed concordance when comparing individual phenotypes. We compared our CaTCH estimates to published estimates from the literature on an individual phenotype basis (Supplementary Note). We found that the correlation for 81 binary and quantitative phenotypes between CaTCH estimates and the published literature was high, r = 0.817 (95% CI: 0.493, 1.14) (Fig. 3a). We also found that 67/81 (82.7%) of phenotypes had overlapping 95% confidence intervals. Of the 81 phenotypes, 49/81 (60.5%) were higher in the published literature.

Discussion

Here we used a large insurance claims dataset to systematically investigate the genetic and environmental contributions in phenotypic variation of 560 phenotypes, including specific environmental risk factors, such as SES, pollution exposure, and climate. Furthermore, we provide estimates of the contributions of genetics and environment in aggregate health cost and comorbidity burden, which are important for both biological research and policy implementation. We also quantified the contribution of one's genetic code and aspects of one's zip code (SES, climate, and air pollution) on the same scale of phenotypic variation for 551 disease-related phenotypes by linking to external geographic databases.

A notable strength of our study was the creation of a large twin cohort. To the best of our knowledge, we amassed the largest twin cohort in the United States that is reflective of household, geographic, and medical-service-based variation of the employed US population. The largest known US twin registries are the Mid-Atlantic Twin Registry (28,000 pairs) and Michigan State Twin Study (15,924). The largest international twin registries are from Sweden (97,000) and Denmark (85,000)[13]. Our twin cohort is comparable in size to these large international twin registries. However, unlike some of these registries, we lack zygosity status for these twin pairs. Furthermore, because we are using insurance claims data, our claims datasets contained the full transactional history between all medical providers and the insurance company for a particular patient. This includes all International Classification of Disease (ICD) 9/10 billing codes sent from the medical provider to the insurance company to be reimbursed. We claim that this provides a comprehensive view into a patient's medical history. In contrast, electronic medical records, because they are a record of the medical examination process, may have deeper phenotypic information (for example, laboratory notes, radiology reports and X-ray images), but will have an incomplete medical history if the patient sees multiple medical providers.

Twin designs have lower sample size than other family-based designs, but are better powered to estimate heritability[14]. However, leveraging the family-based design in a claims-based cohort is not without disadvantages. First, a common issue in insurance data includes a limited observational time window to ascertain phenotype. This can lead to ascertainment bias in phenotypes when siblings are of different ages. This is further exacerbated with analysis including parents and children where, as a result of age of onset,

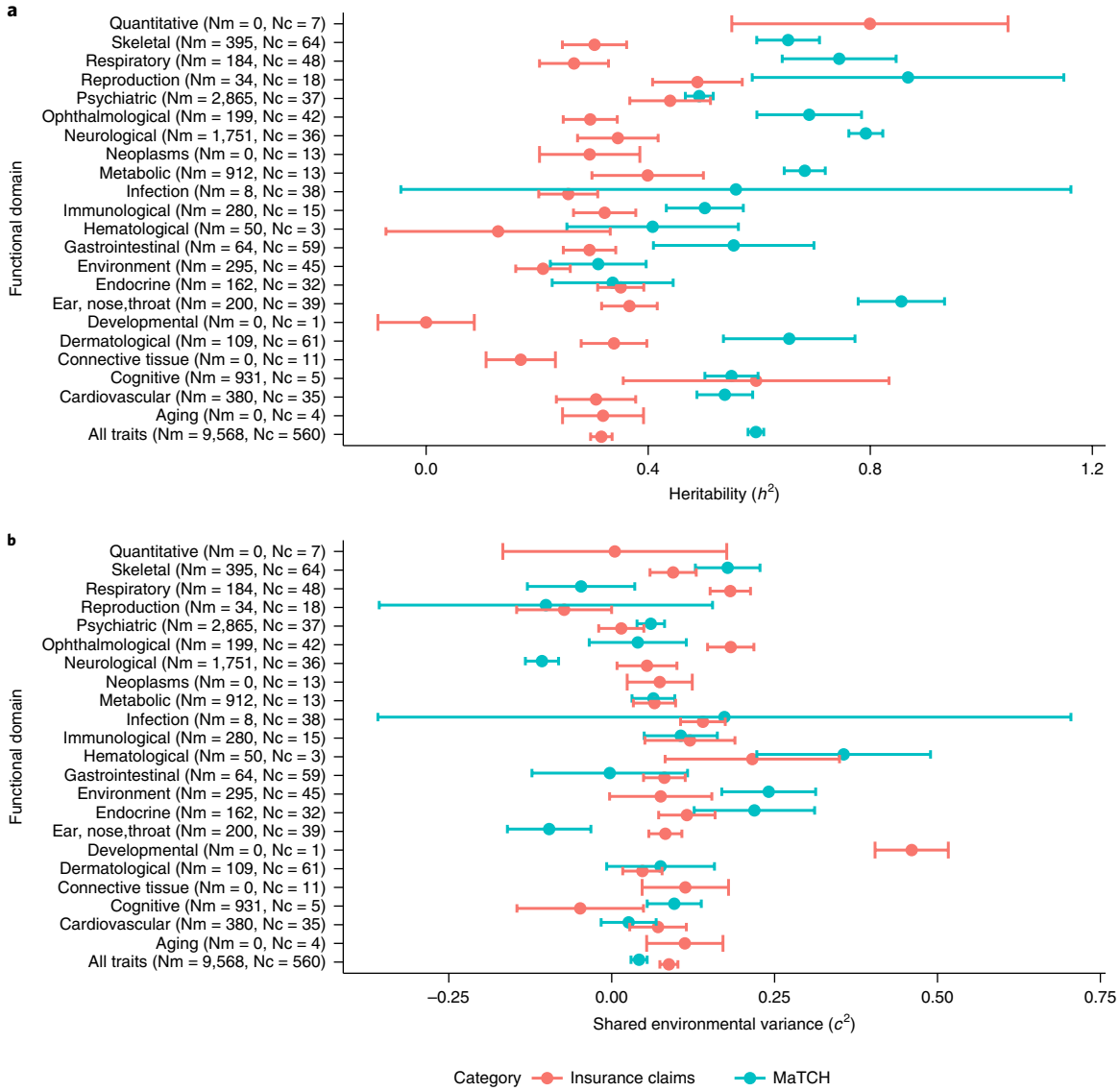

Fig. 4 | Comparison of h^2/c^2 estimates from 56,396 twin pairs among 560 phenotypes in CaTCH to 5,169,880 twin pairs among 9,568 phenotypes in MaTCH (Supplementary Table 1). a, Meta-analytic h^2 estimates for all phenotypes and functional domains between CaTCH and MaTCH; error bars represent 95% CI. Red values are the numbers of CaTCH phenotypes in each functional domain, and blue values are the numbers of MaTCH phenotypes with twin correlation values within each functional domain. **b**, Meta-analytic c^2 estimates for all phenotypes and functional domains between MaTCH and CaTCH; error bars represent 95% CI. Red values are the numbers of CaTCH phenotypes in each functional domain, and blue values are the numbers of MaTCH phenotypes with twin correlation values within each functional domain. Each category is annotated with the number of phenotypes in MaTCH (Nm) and the insurance study (Nc).

the same phenotypic code may represent different disease subtypes[4]. Second, in a family design, estimates of h^2 will be biased[15] if all sources of familial environmental variation are unaccounted (for example, spousal correlation and sibling correlation). Recent family-based studies attempted to estimate some of this familial environmental variation[4,5]; however, limitations remain, such as the lack of interpretability of multiple types of 'shared environment.' [4]. In contrast, twin studies have a simpler design, thereby allowing a single parameter (c^2) to account for all shared environment. Third, claims data do not consider that non-biological relationships can also occur when using next of kin information or subscriber relationships. There is a possibility that 'ascertained' nuclear families may contain step-children, adoptions, or half-siblings; however, this can be modeled using Census data and pedigree simulations[4]. By using both the inferred sibling relationship and the fact that they must be born on the same day, we claim that there is a smaller chance of twins being biologically unrelated.

A major component of our analysis was the ability to compare variance components of specific environmental factors with standard measures used in family-based analysis such as heritability and shared environmental variance. We note that each twin pair has the same shared environment, but our analysis attempts to partition phenotypic variance further with several identified shared environmental factors (Methods and equation (6)) that are common among groups of twin pairs. We believe partitioning the shared environment into identified environmental factors (indicators of local SES, air pollution, and climate) is akin to analysis in partitioning heritability among functional annotations[16–18]. We found that variance components resulting from specific environmental factors were significantly lower than h^2 and c^2 overall and within each functional

Table 2 | Quintiles for each environmental variance component

Quintile	Deprivation index (PC1 component)	Number of pairs	AQI scale	Number of pairs	Average temperature (degrees Fahrenheit)	Number of pairs
1	(−7.516, −1.212)	2,652	(10.580, 33.048)	8,397	(26.190, 50.879)	7,282
2	(−1.212, −0.210)	3,892	(33.048, 37.319)	12,211	(50.879, 55.241)	12,948
3	(−0.210, 0.666)	6,098	(37.319, 41.324)	12,420	(55.241, 60.517)	10,777
4	(0.666, 1.915)	10,838	(41.324, 45.602)	11,525	(60.517, 66.437)	7,844
5	(1.915, 9.601)	24,653	(45.602, 62.721)	3,580	(66.437, 81.295)	9,282

domain (Supplementary Fig. 5). Part of the reason could be a result of choices in how to assess exposure of the environmental risk variables for each particular twin as well as choices in discretizing these variables. In our analysis, we selected environmental variables based on an individual's home residence postal code (zip code) versus individual-level exposure data, which may dilute the influence of these variables on phenotypes. We are limited in our ability to answer (1) how many additional measured shared environmental or non-genetic factors contribute to phenotypic variation beyond geocoded variables and (2) our method requires as input discretized environmental factors. Furthermore, environmental factors may also influence phenotypes through prolonged exposure. In our study, we were underpowered to detect this signal given the young age of our cohort. A natural extension of this research includes approaches to consider continuous environmental variables in these novel and large data streams.

Specific environmental factors had little role in variation of most phenotypes, but we found intriguing results for a few phenotypes. The phenotype with largest socioeconomic variance component was morbid obesity ($var_{SES}=0.027$, 95% CI: 0.014, 0.039). For Lyme disease, the variance components of all three environmental risk factors passed FDR significance for the phenotype ($var_{SES}=0.022$, 95% CI: 0.015, 0.028, $var_{AQI}=0.006$, 95% CI: 0.004, 0.009, $var_{SES}=0.028$, 95% CI: 0.023, 0.033). For lead poisoning, var_{temp} was FDR significant ($var_{temp}=0.029$, 95% CI: 0.017, 0.042).

In the United States, predictors of health care cost and chronically ill patients are of particular importance[19]. In a recent analysis[20] of high-cost patients, the researchers emphasized that prediction of high-cost patients is important, yet current prediction methods do not include any family history information. Our twin analysis concludes that 0.59 of variance for average monthly cost is explained by h^2 and c^2.

Compared to the published literature (as reported by MaTCH) the CaTCH cohort was both younger and had a different distribution of phenotypes. First, in MaTCH, monozygotic correlation, dizygotic correlation, heritability, and shared environmental variance were all smaller, on average, for phenotypes ascertained after adolescence[12]. When comparing h^2 estimates on an individual trait basis the correlation was high ($r=0.817$, 95% CI: 0.493, 1.14). A prerequisite to our analysis is selection of phenotypes with a minimum prevalence threshold and removal of phenotypes with high gender imbalance. Second, we were able to estimate genetic and environmental variance in 38 infectious diseases, compared with only eight phenotypes in MaTCH[12]; on the other hand, phenotypes in psychiatric, metabolic, and cognitive domains accounted for 51% of all twin studies analysed in MaTCH[12]. Such differences in both population and phenotypic selection possibly contribute to differences in estimates versus MaTCH (while still maintaining high correlation for h^2 phenotypes on an individual trait basis), but there may be other methodological differences (such as lack of zygosity information) that may contribute to differences. Our procedure provides an opportunity to investigate phenotypes with large c^2, such as lead poisoning and retinopathy of prematurity (see URLs), whereas many twin studies select phenotypes on the basis of a prior belief of a genetic contribution.

Data on patients from health claims lack zygosity information that is typically ascertained in standard twin registries; however, by amassing a large number of non-twin sibling pairs from the same dataset, we found that the opposite sex twin correlation was close to sibling correlations. For our method to be internally valid, we make the following claims. First, we assume that phenotypic correlation of opposite sex twin pairs (r_{twinOS}) is equivalent to dizygotic same sex twin pairs ($r_{twinDZSS}$). Second, we estimate the proportion of same sex twin pairs are monozygotic by assuming opposite sex and same sex dizygotic twin pairs are equally likely (Methods and equation 19). We tested the first claim by interrogating the concordance between same sex and opposite sex sibling correlations. We found that r_{sibSS} and r_{sibOS} were highly correlated ($r=0.978$, 95% CI: 0.974, 0.981), and, on average, r_{sibSS} was slightly higher than r_{sibOS} (average $r_{sibSS}-r_{sibOS}=0.017$) for the 560 phenotypes passing our filtering criterion (Supplementary Note) and for 23.5% of phenotypes $r_{sibSS}-r_{sibOS}$ followed the null distribution. We conclude that, overall, r_{twinOS} is a proxy for $r_{twinDZSS}$. We note that r_{twinOS} was higher than r_{sibOS} and r_{sibSS} for most phenotypes, suggesting increased h^2 and decreased c^2 if r_{sibOS} or r_{sibSS} were substituted for r_{twinOS} for those traits. We claim that high correlation r_{sibSS} and r_{sibOS} is primarily a result of two factors. First, our phenotypic selection procedure eliminated phenotypes with large imbalances of sex-specific prevalence. Second, we added in sex as a covariate ('fixed-effect') to adjust for the mean differences between males and females. If r_{twinOS} were replaced by r_{sibSS}, then for the majority of phenotypes the estimate of h^2 would increase and c^2 would decrease, raising the possibility that the contribution of the environment may change when assessing siblings rather than twins. We also tested the assumption of using Weinberg's Law, and effect of in vitro fertilization had little to no effect on h^2/c^2 estimates (Supplementary Note).

In our analysis, we ascertained twin pairs between the ages of 0 and 24. This selection criterion eliminated our ability to study late-onset diseases such as Parkinson's and Alzheimer's disease. As with any administrative dataset, there may be errors in ascertainment of phenotype; for example, doctors may not be sure whether a child has type 1 diabetes or type 2 diabetes and therefore may bill for both diseases and therefore the individual may be ascertained as having both diseases. Such bias may be reduced by applying phenotyping algorithms (for example, for diabetes[21]) for each phenotype; however, only a limited number of such algorithms exist.

In summary, our results provide a comprehensive picture of the contribution of genetics and the environment to a large number of phenotypes. We also estimated the contribution of specific environmental risk factors in phenotype. Our estimates provide a useful baseline for determining the potential of further genetic and/or epidemiological research for a number of phenotypes of clinical relevance, applicable and complementary to precision medicine efforts, such as All of US[1].

URLs. American Community Survey: https://factfinder.census.gov/; EPA AQI: https://aqs.epa.gov/aqsweb/airdata/download_files.html#AQI; NOAA Monthly Temperature: https://www.ncdc.noaa.

gov/data-access/land-based-station-data; International Society for Twin Registries: http://www.twinstudies.org/information/twinregisters/; ICD 10 Codes: https://www.cdc.gov/nchs/icd/icd10cm.htm; CaTCH web application, http://apps.chiragjpgroup.org/catch/.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability, and associated accession codes are available at https://doi.org/10.1038/s41588-018-0313-7.

Received: 16 April 2018; Accepted: 7 November 2018;
Published online: 14 January 2019

References

1. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).
2. Roberts, N. J. et al. The predictive capacity of personal genome sequencing. *Sci. Transl. Med.* **4**, 133ra58–133ra58 (2012).
3. Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet.* **6**, e1000864 (2010).
4. Wang, K., Gaitsch, H., Poon, H., Cox, N. J. & Rzhetsky, A. Classification of common human diseases derived from shared genetic and environmental determinants. *Nat. Genet.* **49**, 1319–1325 (2017).
5. Polubriaginof, F. C. G. et al. Disease heritability inferred from familial relationships reported in medical records. *Cell* **173**, 1692–1704.e11 (2018).
6. Benyamin, B., Wilson, V., Whalley, L. J., Visscher, P. M. & Deary, I. J. Large, consistent estimates of the heritability of cognitive ability in two entire populations of 11-year-old twins from Scottish mental surveys of 1932 and 1947. *Behav. Genet.* **35**, 525–534 (2005).
7. Graham, G. N. Why your zip code matters more than your genetic code: promoting healthy outcomes from mother to child. *Breastfeed. Med.* **11**, 396–397 (2016).
8. Slade-Sawyer, P. Is health determined by genetic code or zip code? Measuring the health of groups and improving population health. *N. C. Med. J.* **75**, 394–397 (2014).
9. Heckerman, D. et al. Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proc. Natl Acad. Sci. USA* **113**, 7377–7382 (2016).
10. Denny, J. C. et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
11. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. Series B Stat. Methodol.* **64**, 479–498 (2002).
12. Polderman, T. J. C. et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).
13. van Dongen, J., Eline Slagboom, P., Draisma, H. H. M., Martin, N. G. & Boomsma, D. I. The continuing value of twin studies in the omics era. *Nat. Rev. Genet.* **13**, 640–653 (2012).
14. Docherty, A. R. et al. Comparison of twin and extended pedigree designs for obtaining heritability estimates. *Behav. Genet.* **45**, 461–466 (2015).
15. Liu, C. et al. Revisiting heritability accounting for shared environmental effects and maternal inheritance. *Hum. Genet.* **134**, 169–179 (2015).
16. Loh, P.-R. et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
17. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
18. Lee, S. H. et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).
19. Dieleman, J. L. et al. US Spending on personal health care and public health, 1996–2013. *JAMA* **316**, 2627–2646 (2016).
20. McWilliams, J. M. & Schwartz, A. L. Focusing on high-cost patients - the key to addressing high costs? *N. Engl. J. Med.* **376**, 807–809 (2017).
21. Richesson, R. L. et al. A comparison of phenotype definitions for diabetes mellitus. *J. Am. Med. Inform. Assoc.* **20**, e319–e326 (2013).

Acknowledgements

We thank K. Fox of Aetna, Inc., N. Palmer of Harvard Medical School, and I. Kohane of Harvard Medical School for support and providing access to the Aetna Insurance Claims Data. We are grateful to L. O'Connor and A. Price for helpful discussion. This research was supported by the Australian National Health and Medical Research Council (1078037 and 1113400), National Institutes of Health NIEHS (R00ES23504 and R21ES205052), the National Science Foundation (1636870), and the Sylvia & Charles Viertel Charitable Foundation.

Author contributions

All authors contributed extensively to the work presented in this paper. C.M.L., P.M.V., and C.J.P. designed experiments, analysed data, and wrote the manuscript. B.T.T. developed the Shiny App for analysis. B.T.T., A.K.M., and J.Y. contributed to iterative improvement of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0313-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.M.V. or C.J.P.

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

Methods

Study population. We obtained our data from un-identifiable member claims data from Aetna Inc, a national health insurance company. The claims dataset contained the ICD 9/10 billing codes of 44,859,462 members with an Aetna Insurance plan from January 2008 to February 2016 (Supplementary Fig. 6a). This was a nationally representative dataset; 26,713 of 41,739 US mail zip codes have at least 20 members. We extracted a twin and sibling cohort to estimate genetic and environmental contribution in 560 phenotypes (Supplementary Fig. 6e–k). The twin and sibling cohort focused on younger individuals born on or after 1985 because, under current US health care law, they qualified as dependents on their parent's insurance plans (Supplementary Fig. 6b). In all of our analysis we selected members enrolled for at least 36 consecutive months to have a sufficient period of time for the ascertainment of their phenotypes (Supplementary Fig. 6b).

Twin and sibling cohort creation. We created the twin cohort by extracting primary subscribers and their dependents. Specifically, a primary subscriber would add 'dependents' to his/her policy (approximately 26.19% are sole subscribers) and dependent individuals were coded as 'child', 'grandchild', 'spouse', 'domestic partner', 'legal dependent', and 'student'. We ascertained family structure in this dataset using the relationship between the primary subscriber and child dependents (Supplementary Fig. 6f). We restricted family size to, at most, 15 members living in the same zip code in order to reduce the chance multiple families are merged together (average family size is 3.98) (Supplementary Fig. 6e). Once family units were created, we further extracted twins by comparing the birthdate of child subscribers that are linked to the same primary subscriber. We selected families where there is only one twin pair and eliminated children that are part of a triplet or greater because our estimation of h^2 and c^2 assumed twin pairs are independent and not part of an extended pedigree (Supplementary Fig. 6g).

We created a sibling cohort as a basis for comparison to our twin cohort. Like the twin cohort, the sibling cohort utilized dependent information from the primary subscriber in order to determine sibling pairs (Supplementary Fig. 6j). The sibling cohort also included families where there were at most 15 members, individuals must be born on or before 1985, individuals were enrolled as members for at least 36 months, and each individual had at least one ICD9/10 code (Supplementary Fig. 6b–d). The age difference between sibling pairs had to be at least 11 months and no more than 36 months (Supplementary Fig. 6k). Also, for each family, a single sibling pair that meet these conditions was selected at random (Supplementary Fig. 6k).

Comparison of twin cohort to national population. We compare our twin cohort to the general population using American Community Survey (ACS) Census data. In particular, we ascertained all twins that were members, for at least one year, between 2009–2013 and compared with the 2009–2013 ACS estimates. Using Census data, we estimated a measure for SES for each zip code called the deprivation index, a measure used in epidemiological literature[22] (Supplementary Note). The deprivation index is a measure of SES for a zip code based on seven Census variables that were extracted from the 2009–2013 ACS (see URLs). High deprivation index values correspond to higher SES status and vice versa. For all individuals in the 2009–2013 ACS, we estimated their population density (log transform of number of people per square mile) and deprivation index based on their home zip code and compare to the population density and deprivation index of all twins, enrolled between 2009–2013, based on their home zip code. We observed that more twin pairs live in high population density areas compared to the general population (Fig. 1b). The SES status of twin pairs, based on their home zip code, is slightly higher than the general US population (Fig. 1c).

Phenotype ascertainment. The claims dataset contained all ICD version 9/10 (hereafter ICD9/10, respectively) billing and diagnostic codes provided by the healthcare provider to the insurance company (Aetna, Inc.) for transactional purposes while the individual was a subscriber to the health plan. In practice, many ICD9/10 codes may represent the same overarching phenotype, for example, ICD 250.00 represents type 2 diabetes that is controlled, while 250.02 is type 2 diabetes that is uncontrolled. Thus, we used PheWAS code groupings[10]. PheWAS codes are a way of combining ICD9 codes, used for phenotype-wide association studies[10]. Multiple ICD9/10 codes are combined into a single 'phenotype'. Specifically, an individual was identified as positively having a PheWAS phenotype if they had at least one ICD 9/10 code from the PheWAS code grouping, for example, ICD 9 codes 250.00 and 250.02 both mapped to PheWAS code 250.2 type 2 diabetes. For rarer phenotypes, we utilized the groupings found in Blair, Rzhetsky et al.[23] (we will collectively refer to these phenotypes as PheWAS codes). In total, we mapped Aetna subscriber ICD9/10 diagnostic codes to 1,900 PheWAS codes (Supplementary Fig. 6c). PheWAS mappings were originally constructed using ICD9 codes, but the surveillance period for the insurance data spanned the transition from ICD9 to ICD10. In order to accommodate ICD10 codes, we utilized the United States Center for Disease Control and Prevention 2016 General Equivalence Mapping of ICD10 (see URLs) codes to ICD9 and subsequently to PheWAS codes.

For a subset of individuals, the claims dataset provided results of diagnostic clinical laboratory tests (hereafter called 'lab test') conducted during the individual's medical care (Supplementary Fig. 6c). Each lab test was identified by a logical observation identifier name and code[24]. For only the twin cohort, we ascertained all lab tests where twin pairs were measured on the same day. In our analysis we included all laboratory tests where there were at least 2,000 twin pairs that match our criterion. The phenotypes we analysed include common laboratory tests such as low density lipoprotein cholesterol, high density lipoprotein cholesterol, triglycerides, leukocyte counts and hemoglobin counts. If a twin pair had multiple lab tests, then we randomly sampled a single lab test event for analysis.

Out of a total of 1,900 binary phenotypes, we removed phenotypes with low prevalence or where disparity in male and female prevalence was high (Supplementary Fig. 6d) among twin pairs. In particular, for each phenotype, we imposed a filtering criterion where the ratio of male prevalence to female prevalence (or female to male prevalence) among twin pairs must be less than five (Supplementary Fig. 6d). In addition, only phenotypes with a prevalence of at least 0.3% were kept, resulting in phenotypes where at least 338 cases were expected and at least one concordant same sex and opposite sex pair allowing for stable estimation of h^2 and c^2, resulting in 551 binary phenotypes. In the case of the quantitative phenotypes, we analysed laboratory values that had at least 2,000 twin pairs (Supplementary Fig. 6d). For the sibling pairs, we ascertained only the 551 binary phenotypes.

For the twin cohort, in the claims dataset, we utilized an opportunity to derive phenotypes based on aggregate claims, including the total number of PheWAS codes per individual (or comorbidities) and the average monthly cost incurred per individual (hereafter called 'average monthly cost'). The number of PheWAS codes was the number of distinct PheWAS codes ascertained for a patient during the time of surveillance (at least 36 months) and can be thought of as the total number of 'comorbidities' coded for each individual. Average monthly cost was the total claim costs divided by the months that the individual was a member of this insurance company when the costs were incurred.

Specific environmental risk factors. For each twin pair we ascertained their home zip code and linked to Census data deprivation index, daily air quality index data, and monthly average temperature data. The deprivation index is a composite score of SES for a zip code based on seven variables from the 2009–2013 ACS (Supplementary Note). The Environmental Protection Agency used the air quality index (AQI) to summarize air pollution level in a particular location. The AQI has a range between 0 and 500. An AQI value between 0–50 is considered good air quality, 50–100 is moderate air quality and above 100 is considered unhealthy air quality. We downloaded all daily county-level AQI data provided by the EPA (see URLs) and estimated the median AQI level exposure for each twin pair based on the twin pairs dates of enrollment and closest county to their zip code (maximum distance of 30 km) (Fig. 1d). We also ascertained all monthly average temperature data from sensors located throughout the United States from the National Atmospheric and Oceanic Administration (NOAA) (see URLs). For each twin pair, we found the closest NOAA sensor to their home zip code and extracted all monthly average temperature data based on their months of enrollment within the insurance claims dataset, then estimated the median monthly average temperature based on those values (Fig. 1e). This linkage provided, for each twin pair, a quantitative measurement for median family income, median AQI and median monthly average temperature based on their home zip code. The quantitative value for each environmental risk factor was binned into quintiles based on the distribution of the quantitative value among the general US population (see Table 2 for the ranges and number of twin pairs in each quintile).

Variance component model for twin data. Estimation of heritability (h^2), and shared environmental variance (c^2) all rely on the estimation of various variance component parameters on the observed scale. Following the convention in Visscher et al.[25], the variance component model can be written:

$$y = X\beta + \sum_{i=1}^{k} u_i + e \quad (1)$$

where $y = 1$ for individuals who had a PheWAS code and $y = 0$ for individuals who did not have a PheWAS code for a binary phenotype, y is a real-valued inverse normal rank transformation of the lab test or utilization trait values[26] for quantitative phenotypes, $X\beta$ are fixed effects that were sex, months of enrollment and age (average age during surveillance for PheWAS phenotypes and derived quantitative phenotypes or age of test for lab tests) in our model. The terms $u_i \sim N(0, V_i)$ were random effects used to estimate all variance components for this analysis and e is the error term.

In the twin cohort, we used the variance component model to estimate h^2, c^2 and environmental risk random effects. See Supplementary Note for estimation of opposite sex and same sex sibling correlation (Supplementary Note). All twin estimates relied on the model

$$y = X\beta + u_{\text{pair}} + u_{\text{extraSS}} + e \quad (2)$$

where $\text{var}(y) = V_{\text{pair}} + V_{\text{extraSS}} + V_e$. The random effect u_{pair} is common to a pair of both opposite sex and same sex twin pairs, while u_{extraSS} is common to a pair of same sex

pairs but different for opposite sex pairs, thus the covariance between individuals i and j in a pair is $\mathrm{cov}(y_i, y_j) = V_{\mathrm{pair}}$ for opposite sex pairs and $\mathrm{cov}(y_i, y_j) = V_{\mathrm{pair}} + V_{\mathrm{extraSS}}$ for same sex pairs. Same sex and opposite sex variance components were estimated as follows:

$$V_{\mathrm{twinSS}} = V_{\mathrm{pair}} + V_{\mathrm{extraSS}} \tag{3}$$

$$V_{\mathrm{twinOS}} = V_{\mathrm{pair}} \tag{4}$$

$$V_{\mathrm{tot}} = V_{\mathrm{pair}} + V_{\mathrm{extraSS}} + V_{\mathrm{res}} \tag{5}$$

This model was extended to include environmental risk random effects u_{SES}, u_{AQI} and u_{temp} based on the quintiles (Table 2) for each environmental risk factor, written as follows:

$$y = X\beta + u_{\mathrm{pair}} + u_{\mathrm{extraSS}} + u_{\mathrm{SES}} + u_{\mathrm{AQI}} + u_{\mathrm{temp}} + e \tag{6}$$

The random effects u_{pair} and u_{extraSS} are the same as in equation (2), while the random effects u_{SES}, u_{AQI} and u_{temp} will be common to all individuals belonging to the same deprivation index, AQI or temperature quantile bin, respectively.

Estimation of twin same sex and opposite sex correlation. We used variance components V_{twinSS} and V_{twinOS} to estimate h^2 and c^2 by first transforming them into correlation on the observed scale:

$$r_{\mathrm{twinSS01}} = \frac{V_{\mathrm{twinSS}}}{V_{\mathrm{tot}}} \tag{7}$$

$$r_{\mathrm{twinOS01}} = \frac{V_{\mathrm{twinOS}}}{V_{\mathrm{tot}}} \tag{8}$$

Conversion of binary phenotypes to liability scale. In the case of quantitative (real-valued) phenotypes, we used correlations r_{twinSS01} and r_{twinOS01} on the observed scale to estimate h^2 and c^2, but in the case of binary phenotypes we transformed these correlations onto the liability scale. The transformation of correlation from the observed scale to the liability scale was estimated as follows (opposite sex formulas are same as same sex)[27]:

$$T = \Phi^{-1}(1-K) \tag{9}$$

$$z = \Phi(T) \tag{10}$$

$$i = \frac{z}{K} \tag{11}$$

$$Eb_{\mathrm{twinSS}} = K + \frac{V_{\mathrm{twinSS}}}{K} \tag{12}$$

$$T_{\mathrm{twinSS}} = \Phi^{-1}(1-Eb_{\mathrm{twinSS}}) \tag{13}$$

$$r_{\mathrm{twinSS}} = \frac{(T-T_{\mathrm{twinSS}})\sqrt{1-(T^2-T^2_{\mathrm{twinSS}})\left(1-\frac{T}{i}\right)}}{i+T^2_{\mathrm{twinSS}}(i-T)} \tag{14}$$

K is the population prevalence for the phenotype (estimated from filtered population) and Φ was the standard normal distribution. The formulas for r_{twinSS} and r_{twinOS} accounted for the reduction of variance expected from the relatives of proband compared to the general population[27].

Similarly, the variance components for environmental risk factors ($\mathrm{var}_{\mathrm{SES}}$, $\mathrm{var}_{\mathrm{AQI}}$ or $\mathrm{var}_{\mathrm{temp}}$) on the liability scale were estimated as follows ($\mathrm{var}_{\mathrm{env}}$ for $\mathrm{var}_{\mathrm{env}} = \mathrm{var}_{\mathrm{SES}}$, $\mathrm{var}_{\mathrm{AQI}}$ and $\mathrm{var}_{\mathrm{temp}}$):

$$Eb_{\mathrm{env}} = K + \frac{V_{\mathrm{env}}}{K} \tag{15}$$

$$T_{\mathrm{env}} = \Phi^{-1}(1-Eb_{\mathrm{env}}) \tag{16}$$

$$\mathrm{var}_{\mathrm{env}} = \frac{(T-T_{\mathrm{env}})\sqrt{1-(T^2-T^2_{\mathrm{env}})\left(1-\frac{T}{i}\right)}}{i+T^2_{\mathrm{env}}(i-T)} \tag{17}$$

Estimation of heritability and shared environmental variance. In traditional twin studies, where zygosity of twins were known, the h^2 and c^2 of a phenotype were calculated using the monozygotic (MZ) twin correlation r_{twinMZ} and dizygotic (DZ) same sex twin correlation r_{twinDZSS} as follows[28]:

$$h^2 = 2(r_{\mathrm{twinMZ}} - r_{\mathrm{twinDZSS}}) \tag{18}$$

$$c^2 = 2r_{\mathrm{twinDZSS}} - r_{\mathrm{twinMZ}} \tag{19}$$

In a health administration dataset, the zygosity status of twins is not known. However, opposite sex twin pairs are dizygotic and same sex twin pairs are a mixture of monozygotic and dizygotic pairs. Assuming the probability of a dizygotic twin pair being same sex is 50% (Weinberg's Rule[29]), we estimated the probability (p) of a pair being monozygotic given they are same sex is calculated as follows[6,30,31]:

$$p(\mathrm{MZ}) = 1 - 2p(\mathrm{OS}) = 1 - 2\frac{N_{\mathrm{OS}}}{N_{\mathrm{SS}}} \tag{19}$$

$$p(\mathrm{SS}) = \frac{N_{\mathrm{SS}}}{N_{\mathrm{all}}} \tag{20}$$

$$p = p(\mathrm{MZ}|\mathrm{SS}) = \frac{p(\mathrm{MZ})}{p(\mathrm{SS})} \tag{21}$$

where N_{all} was the total number of twin pairs, N_{OS} was the number of opposite sex pairs and N_{SS} was the number of same sex pairs. Assuming r_{twinOS} was equal to r_{twinDZSS} and r_{twinSS} was a mixture of r_{twinDZSS} and r_{twinMZ} then h^2 and c^2 were estimated as follows:

$$r_{\mathrm{twinOS}} = r_{\mathrm{twinDZSS}} \tag{22}$$

$$r_{\mathrm{twinSS}} = pr_{\mathrm{twinMZ}} + (1-p)r_{\mathrm{twinDZSS}} \tag{23}$$

$$h^2 = \frac{2}{p}(r_{\mathrm{twinSS}} - r_{\mathrm{twinOS}}) \tag{24}$$

$$c^2 = \frac{(p+1)r_{\mathrm{twinOS}} - r_{\mathrm{twinSS}}}{p} \tag{25}$$

We estimated standard errors for r_{twinOS}, r_{twinSS}, h^2, c^2, $\mathrm{var}_{\mathrm{SES}}$, $\mathrm{var}_{\mathrm{AQI}}$ and $\mathrm{var}_{\mathrm{temp}}$ via bootstrap resampling (500 samples). In the analysis of binary phenotypes and derived quantitative phenotypes, which use the full twin cohort, the parameter p was 0.42. We estimated the parameter p for quantitative phenotypes, using equation (21), based on the subset of twins that had that particular quantitative phenotype (Supplementary Note). The p estimates for quantitative phenotypes ranged from 0.513 to 0.572.

Multiple comparisons. For all statistics (variance components h^2, c^2, $\mathrm{var}_{\mathrm{SES}}$, $\mathrm{var}_{\mathrm{AQI}}$ and $\mathrm{var}_{\mathrm{temp}}$ and fixed effects β_{age} and β_{sex}) we estimated P values using a two-tail z-test statistic and we accounted for multiple hypothesis testing by controlling by estimating the FDR. In particular, we used the Benjamini–Yekutieli[32] method to estimate the FDR rate that assumes dependencies between phenotypes. We estimated FDR adjusted P values for all statistics and report the number of phenotypes, for each statistic, which achieved FDR < 5%.

We fit all random effects models with the 'lme4' package in R[33]. We wrote our own bootstrapping procedure in order to estimate standard errors for all statistics presented in this paper. We used the p.adjust function in the base stats R package[34] for FDR correction.

Matching of PheWAS codes to functional domains from MaTCH. We sought to compare how h^2 and c^2 estimates compared to the published literature. To enhance comparison, we downloaded h^2 and c^2 estimates from a large and recent meta-analysis of twin studies[12]. We mapped PheWAS codes into functional domains as determined by the MaTCH study[12]. Each functional domain constituted a subset of chapters and subchapter levels from either the International Classification of Functioning, Disability and Health or International Statistical Classification of Diseases and Related Health Problems (ICD-10). In the claims dataset, we mapped each PheWAS code to their constituent ICD9 code and then mapped again to the corresponding ICD10 chapters and subchapters. If the associated chapter or subchapter from a PheWAS code overlapped with a functional domain then we considered it part of the domain. We estimated the mean h^2 and c^2 for each domain with an inverse-variance weighting estimate. We also estimated the number of

phenotypes that follow a model due to additive genetic variance and not non-additive genetics (including dominance) or shared environmental variance, which was estimated by the number of phenotypes that follow $2r_{twinDZSS} = r_{twinMZ}$. This was equivalent to the number of phenotypes that follow the null hypothesis (pi_0 statistic[11]) $c^2 = 0$, which was directly estimated in our study.

Overall and functional domain values of h^2 and c^2 were calculated with the 'metafor'[35] R package by using the DerSimonian–Laird[36] estimator to calculate estimates and standard errors. The pi_0 statistic was estimated using the 'qvalue'[37] R package.

Comparison of h^2 estimates to published literature. In our analysis, we compared h^2 estimates from the published literature to h^2 estimates from CaTCH (Supplementary Note). The correlation between CaTCH h^2 estimates and published h^2 estimates used a correlation estimator[37] that also incorporated standard errors. We used jackknife resampling in order to estimate the standard error for this estimator, as suggested by the authors of this method[37].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data Availability

The data that support the findings of this study are available from Aetna Insurance, but restrictions apply to the availability of these data, which were used under licence for the current study, and so are not publicly available. Please contact N. Palmer (nathan_palmer@hms.harvard.edu) for inquiries about the Aetna dataset. Summary data are, however, available from the authors upon reasonable request and with permission of Aetna Insurance. Code for analysis, generation of figures and figure files is available at https://github.com/cmlakhan/twinInsurance.

References

22. Krieger, N. et al. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: the public health disparities geocoding project (US). *J. Epidemiol. Community Health* **57**, 186–199 (2003).
23. Blair, D. R. et al. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. *Cell* **155**, 70–80 (2013).
24. Huff, S. M. et al. Development of the logical observation identifier names and codes (LOINC) vocabulary. *J. Am. Med. Inform. Assoc.* **5**, 276–292 (1998).
25. Visscher, P. M., Benyamin, B. & White, I. The use of linear mixed models to estimate variance components from data on twin pairs by maximum likelihood. *Twin. Res.* **7**, 670–674 (2004).
26. Beasley, T. M., Erickson, S. & Allison, D. B. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behav. Genet.* **39**, 580–595 (2009).
27. Reich, T., James, J. W. & Morris, C. A. The use of multiple thresholds in determining the mode of transmission of semi-continuous traits. *Ann. Hum. Genet.* **36**, 163–184 (1972).
28. Falconer, D. S. & Mackay, T. C. *Introduction to Quantitative Genetics* (John Wiley & Sons. Inc., New York,, 1989).
29. Weinberg, W. Beiträge zur Physiologie und Pathologie der Mehrlingsgeburten beim Menschen. *Pflugers Arch. Gesamte Physiol. Menschen Tiere* **88**, 346–430 (1901).
30. Neale, M. C. A finite mixture distribution model for data collected from twins. *Twin. Res.* **6**, 235–239 (2003).
31. Scarr-Salapatek, S. Race, social class, and IQ. *Science* **174**, 1285–1295 (1971).
32. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
33. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
34. R. C. Team R: A language and environment for statistical computing (R Foundation for Statistical Computing, 2014).
35. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36**, 1–48 (2010).
36. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177–188 (1986).
37. Qi, T. et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.* **9**, 2282 (2018).

nature research

Corresponding author(s): Chirag J Patel and Peter M Visscher

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	All claims data was stored in a Microsoft SQL Server 2014 database. We wrote scripts consisting of SQL queries to extract data from raw insurance data into a form suitable for analysis. SQL scripts are not publicly available due to sensitivity in exposing proprietary insurance information, but can be made available to reviewers upon request.
Data analysis	We performed all data analysis using R version 3.3.2 and libraries therein. In particular data processing used tools from the tidyverse (1.1.1) library and random effects modeling used the lme4 library (version 1.1-14). R scripts used to perform our analysis can be found in our github repo (https://github.com/cmlakhan/twinInsurance).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study are available from Aetna Insurance but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Summary data are however available from the authors upon reasonable request and with permission of Aetna Insurance.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We are repurposing an existing dataset therefore our sample size was determined based on all individuals that fit our filtering criterion. Our filtering criterion is described in the Online Methods section of our manuscript.
Data exclusions	We repurposed an existing dataset which was not meant for twin studies. Therefore, we applied multiple filtering criterion in order to best ascertain twin pairs with a long surveillance period in order to ascertain phenotypes. We describe our filtering process in detail in the online methods section.
Replication	N/A
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging