Evaluating the contribution of genetic and familial shared environment to common disease using the UK Biobank

Title:

Evaluating the contribution of genetic and familial shared environment to common disease using the UK Biobank

Authors:

María Muñoz[1], Ricardo Pong-Wong[1], Oriol Canela-Xandri[1], Konrad Rawlik[1], Chris S. Haley[1,2], Albert Tenesa[1,2,3]

Affiliations:

[1]The Roslin Institute, Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, UK.

[2] MRC Human Genetics Unit at the MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road South, Edinburgh, EH4 2XU, UK

[3] Corresponding author

Dr Albert Tenesa

The Roslin Institute

The University of Edinburgh

Easter Bush

Roslin, Midlothian

EH25 9RG

UK

Tel: 0044 (0)131 651 9100

Fax: 0044 (0)131 651 9220

Email: Albert.Tenesa@ed.ac.uk

ABSTRACT

Genome-wide association studies have detected many loci underlying susceptibility to disease, but most of the genetic factors that contribute

to disease susceptibility remain unknown. Here we provide evidence that part of the missing heritability can be explained by an overestimation of heritability. We estimated the heritability of twelve complex human diseases using family history of disease in 1,555,906 white European individuals from the UK Biobank. Estimates using simple family-based statistical models were inflated on average by ~47% comparing with those from Structural Equation Models (SEM) that specifically accounted for shared familial environmental factors. In addition, heritabilities using SNP data explained an average of 44.2% of the simple family-based estimates across diseases and an average of 57.3% of SEM estimated heritability and accounted for almost all of the SEM heritability for hypertension. Our results show that both genetics and familial environment make substantial contributions to familial clustering of disease.

INTRODUCTION

The causation of most common human diseases is complex, being influenced by a combination of genetic and environmental factors[1]. The development of genome-wide association studies (GWAS) has allowed the detection of many genetic variants associated with these diseases. However, these variants only explain a fraction of the heritability estimated in previous family-based studies and hence there is a "missing heritability" that remains unidentified[2]. One possible explanation for this missing heritability is that previous heritability estimates could be inflated because family environmental effects were not specified in the model or because they could not be estimated due to the study design[3]. Furthermore, comparisons of heritability explained by SNPs identified through GWAS or the hidden heritability estimated from genome-wide arrays (that is, the SNP heritability which captures the contribution of common variants including those not yet detected as genome-wide significant due to lack of power) with published estimates of heritability possess some important challenges. For instance, the populations from which family-based heritability

estimates were obtained may differ from those used in the GWAS studies in definition or prevalence of disease or genetic background. These, and other factors[3], make assessments of heritability estimates for disease from familial and GWAS studies difficult and in some instances inappropriate.

The objective of the current study was to estimate the heritability of twelve complex human diseases using self-reported personal and family history of disease in 1,555,906 white European participants and relatives from the UK Biobank, which comprise over 2% of the UK population.

RESULTS

Data overview and Relative Risks

The UK Biobank contains disease and trait data, as well as biological samples collected from around 500,000 participants and has as its main objective to identify ways of improving the prevention, diagnosis and treatment of complex diseases[4]. UK Biobank participants were measured for multiple traits and questioned about their lifestyle, environmental risk factors and medical history and gave their informed consent following strict protocols[5]. Here we use information from the family disease history reported by participants to estimate the heritability and the environmental contributions to the liability of twelve broadly defined complex diseases: heart disease, stroke, chronic bronchitis, hypertension, diabetes, Alzheimer's disease, Parkinson's disease, severe depression and lung, bowel, prostate and breast cancers (Supplementary Table 1). Accuracy of self-reported health status was assessed and is discussed in the supplementary information (Supplementary note and Supplementary Tables 2 and 3).

Disease prevalence was higher among men than among women for all diseases except for severe depression, which was more prevalent among women (Supplementary Table 4). Generally, disease prevalence was higher among the parents of the participants than among the participants and their siblings, suggesting an age-related increase in disease liability. The relative risks (RR) of parents (RR_{PO}) and siblings (RR_{SIB}) of ill individuals participating in

UK Biobank were estimated for each disease. In addition, the relative risk for partners of affected individuals was estimated using information from the parents of the participants (RR_{PAR}). All the relative risks were significantly larger than one (Supplementary Figure 1). Overall, the relative risks for the estimates of RR_{PO} and RR_{SIB} that combined information from blood and adopted relatives were higher than those for $RR_{PAR,}$ except for hypertension and lung cancer. These estimates of relative risks suggest that combinations of both genetics and shared environmental risk factors contribute to the causation of these diseases (Supplementary Figure 1).

Heritability Estimates using Falconer's Method

We estimated heritability values (h^2) from either the correlations or regression coefficients (*b*) of the first-degree family pairs: parents-offspring (participants), siblings-participants and parents-siblings of participants (to provide h^2_{PO}, h^2_{SIB}, and h^2_{PSIB}, respectively) following Falconer's Method (Methods). Correlations or regression coefficients using information of adoptive parent-offspring (b_{APO}) and adoptive sibling (b_{ASIB}) pairs, and parents of participants (partners, b_{PAR}) were also calculated. Estimates among concordant and discordant gender pairs were calculated using a method that takes into account differences between sexes[1], then these estimates were combined using a weighted mean of *b* across all gender pairs. Across generation differences in disease prevalence were taken into account using a control population of the same age for comparison[1]. Genetic correlations between genders were close to one, but tended to be lower than one (Supplementary Table 5).

All heritability estimates from first-degree family pairs were significantly different from zero (Table 1). The highest h^2_{PO} value was noted for depression (0.491±0.007) whereas the highest h^2_{SIB} was observed for prostate cancer (0.707±0.062). Estimates of h^2_{PO} were significantly lower than those of h^2_{SIB} for heart disease, stroke, hypertension, diabetes, and prostate and breast cancers, suggesting the existence of non-additive genetic effects or a greater environmental similarity between siblings than between parents and their children. The highest value of the regression from adoptive parent-offspring

pairs, b_{APO}, was observed for severe depression (0.250±0.036) suggesting an important influence of shared family environmental effects on this disease. The adoptive parent-offspring regression, although much smaller than for depression, was also significantly greater than zero for heart disease, bronchitis and breast cancer. Hypertension had a high value for the correlation between partners, $b_{PAR,}$ (0.203±0.002) and a low value for b_{APO} (0.035±0.021) indicating the importance of environmental effects shared by partners but that are not shared between parents and their offspring, and/or positive assortative mating for hypertension or a trait or combination of traits highly correlated with hypertension.

Significant positive regression or correlation coefficients from adopted pairs and partners (e.g. parents of participants) suggest the potential existence of various environmental effects shared by family members. Hence estimates of heritability obtained using only blood relatives or from models that do not account for the full complexity of shared environmental effects may be inflated (Supplementary Table 6)[6-13].

Heritability Estimates using Structural Equation Modelling

Heritabilities estimated from SEM were in general lower than those estimated using Falconer's method, with significant family environmental effects detected for all the diseases except for Parkinson's disease (Table 2, Supplementary Table 7). Although for most diseases, genetic effects were the major attributable contribution to disease liability, for hypertension the sum of the effects due to shared familial environment was more important than genetic effects (A = 0.28 and C+S+P = 0.33). The estimated partner effect (P = 0.13) for hypertension and the common family effect (C = 0.15) for depression were high. High values of P inform about shared environment among partners or perhaps the presence of assortative mating. The physiological nature of hypertension mitigates against the possibility of assortative mating and it seems more likely that the high estimate for P is due to environmental factors shared by partners such as diet. However we cannot conclusively differentiate among these possibilities without more information such as the length of cohabitation[14]. The relatively large estimate of the common family effect for depression (C=0.15) would

account for approximately half the correlation in the liability for depression between first degree relatives (as the expected correlation = A/2 + C) and would be important to consider in future studies of depression. Similarly our estimates suggest that at least a half of the correlation in disease liability between siblings is due to the combined effects of common family (C) and sibling (S) environment for heart disease, hypertension and lung cancer.

Simulations

To test the performance of our analytical methods we simulated data for the twelve different diseases using the genetic and environmental contributions to liability estimated under the full model for each disease and the corresponding values of prevalence for fathers, mothers, participants and siblings (Supplementary Table 8). We analyzed ten replicated simulations for each disease to estimate the liability components. The means of liability components of the ten replicates were similar to those used to perform the simulations (Supplementary Tables 8 and 9). Performing model comparison within each replicate (Methods), recovered the model used to simulate data in more than 50% of the replicates for 4 of the 12 diseases (heart disease, hypertension, severe depression and prostate cancer) (Supplementary Table 10). However, even for the instances where the true generating model was not recovered, the means of genetic parameters across replicates were similar to those used to simulate data (Supplementary Table 11). Fitting an AE model ignoring familial environment to the simulated data yielded an overestimation of the heritability for all diseases (Supplementary Table 12).

Heritability using SNPs

We obtained SNP heritability estimates using 525,242 SNPs in the genotyped subsample of 114,264 unrelated individuals for those diseases with prevalence higher than 0.50% (Methods). The SNPs explained an average of 44.2% of the Falconer's method estimates, 44.0% of the SEM family–based heritability estimates using the AE model (Omitting family environmental factors - Supplementary Table 13) and 57.3% of the SEM family–based heritability

estimates under the most parsimonious adequate model including family environmental factors, respectively, across diseases. The SNP heritability explained ~100% of the SEM heritability estimate for hypertension (Figure 1, Table 3), which suggests that, for this high-prevalence disease where we could model a large number of familial environmental factors, there might be little or no missing heritability. Conclusions from SNP heritability estimates were similar when SNPs were split into common and rare minor allele frequency (MAF) groups and the joint heritabilities of these two groups were estimated (Supplementary Table 14). However, as previously reported by Mancuso *et al*[15] and Yang *et al*[16] and) these estimates were generally slightly lower than estimates based on a single variance component of common and rare variants.

SNP heritability estimates from self-reported and medical records (Supplementary Table 15) were not significantly different from each other, supporting the usefulness of the self-reported records. This was further confirmed by the similarity in the number of published GWAS hits with significant associations in the UK Biobank data using the self-reported definition of disease or the definition of disease from medical records (Supplementary Table 16).

DISCUSION

In the current study, we estimated the heritabilities of twelve diseases from family-based data using a model which does not take into account environmental factors shared by the family members (Falconer's method) and a SEM method which enables joint estimation of these environmental factors and genetic factors. For most diseases, we obtained lower heritability values with the SEM method than with Falconer's method associated with significant shared environmental effects. Therefore, the heritability estimates using SNPs were closer to SEM family-based heritability values than to those from Falconer's method. Indeed for hypertension the heritability estimates using SNPs was similar to the SEM family-based heritability.

Recently, Yang *et al*[16] have used information from simulated and observed data and analysis of high density imputed data to conclude that there is limited evidence for missing heritability for height and BMI once potential overestimation of heritability in family-based studies is taken into account. Zaitlen *et al*[17] studied twenty three traits in the Icelandic population and suggested that most of the "missing heritability" is likely due to rare variants not included in the genotyping array but also reported that the excess correlation among close relatives was mostly accounted for by shared environment. Finally, Liu *et al*[18] have also shown that models accounting for a diverse source of shared environmental effects should be tested to avoid bias in heritability estimation for a number of quantitative traits. In agreement with Zaitlen *et al*[17], our very large study provides evidence that part of the missing heritability may be due to previous inflated heritability estimates and demonstrates this for important binary disease traits.

This study was based on a large cohort from the UK population, allowing us to estimate heritability with much narrower confidence intervals than in previous studies. In addition, models accounting for different environmental components shared by family members could be implemented due to the information available for different first-degree blood and adoptive relatives. The twelve diseases analyzed in this large cohort of individuals show significant but moderate values of heritability and an important impact of shared familial environmental effects and support the case for combining these factors with genetic marker information in order to improve the performance of disease-risk prediction methods[19,20]. Our results are very relevant when assessing the potential for the development of personalized medicine, providing realistic expectations of the value of genetic testing. In addition, demonstration of the importance of environmental risk factors that contribute to the aggregation of disease within families motivates research to identify and moderate these factors.

URLs

UK Biobank, http://www.ukbiobank.ac.uk/; plink, https://www.cog-genomics.org/plink2; ARCHER UK National Supercomputing Service, http://www.archer.ac.uk; genotyping procedure and genotype calling protocols of the UK Biobank, http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=155580; UK Biobank internal QC procedures, http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=155580; DISSECT, https://www.dissect.ed.ac.uk; GWAS catalogue, https://www.genome.gov/gwastudies/

ACKNOWLEDGEMENTS

This research has been conducted using the UK Biobank Resource, and funded by the Roslin Institute Strategic Programme Grant from the BBSRC (BB/J004235/1). CSH and AT also acknowledge funding from the Medical Research Council. We thank Ian White for his helpful comments. Heritability estimates using SNPs were performed using the ARCHER UK National Supercomputing Service.

AUTHOR CONTRIBUTIONS

AT and CSH conceived and designed the study. MM and AT performed the statistical analysis. OCX and KR carried out the SNP filtering and QC. MM, CSH and AT wrote the manuscript. RPW performed the simulations and contributed ideas and quantitative genetics expertise. All authors read and approved the manuscript.

COMPETING INTEREST STATEMENT

The authors declare no competing financial interests.

REFERENCES

1. Falconer, D.S. Inheritance of Liability to Certain Diseases Estimated from Incidence among Relatives. *Annals of Human Genetics* **29**, 51-& (1965).
2. Manolio, T.A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
3. Tenesa, A. & Haley, C.S. The heritability of human disease: estimation, uses and abuses. *Nat Rev Genet* **14**, 139-49 (2013).
4. Allen, N. et al. UK Biobank: Current status and what it means for epidemiology. *Health Policy Technol* **1**, 123-126 (2012).
5. Centre, U.B.C. UK Biobank: Protocol for a large-scale prospective epidemiological resource. (Protocol No: UKBB-PROT-09-06 (Main Phase), 2007).
6. Kaprio, J. et al. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia* **35**, 1060-7 (1992).
7. Lichtenstein, P. et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* **343**, 78-85 (2000).
8. Iliadou, A. et al. Repeated blood pressure measurements in a sample of Swedish twins: heritabilities and associations with polymorphisms in the renin-angiotensin-aldosterone system. *J Hypertens* **20**, 1543-50 (2002).
9. Gatz, M. et al. Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry* **63**, 168-74 (2006).
10. Zdravkovic, S., Wienke, A., Pedersen, N.L. & de Faire, U. Genetic influences on angina pectoris and its impact on coronary heart disease. *Eur J Hum Genet* **15**, 872-7 (2007).
11. Hallberg, J. et al. Interaction between smoking and genetic factors in the development of chronic bronchitis. *Am J Respir Crit Care Med* **177**, 486-90 (2008).
12. Korja, M. et al. Genetic epidemiology of spontaneous subarachnoid hemorrhage: Nordic Twin Study. *Stroke* **41**, 2458-62 (2010).

13. 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).
14. Annest, J.L., Sing, C.F., Biron, P. & Mongeau, J.G. Familial aggregation of blood pressure and weight in adoptive families. II. Estimation of the relative contributions of genetic and common environmental factors to blood pressure correlations between family members. *Am J Epidemiol* **110**, 492-503 (1979).
15. Mancuso, N. et al. The contribution of rare variation to prostate cancer heritability. *Nature Genetics* **48**, 30-5 (2016).
16. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* (2015).
17. Zaitlen, N. et al. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genetics* **9**, e1003520 (2013).
18. Liu, C. et al. Revisiting heritability accounting for shared environmental effects and maternal inheritance. *Hum Genet* **134**, 169-79 (2015).
19. Aschard, H., Vilhjalmsson, B.J., Joshi, A.D., Price, A.L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *American Journal of Human Genetics* **96**, 329-39 (2015).
20. Tada, H. et al. Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history. *European Heart Journal* **37**, 561-7 (2016).
21. Risch, N. Linkage strategies for genetically complex traits. I. Multilocus models. *American Journal of Human Genetics* **46**, 222-8 (1990).
22. Falconer, D.S. & Mackay, T.F. *Introduction to Quantitative Genetics*, (1996).
23. Dempster, E.R. & Lerner, I.M. Heritability of Threshold Characters. *Genetics* **35**, 212-36 (1950).
24. Boker, S. et al. OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika* **76**, 306-317 (2011).
25. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

26. Canela-Xandri, O., Law, A., Gray, A., Woolliams, J.A. & Tenesa, A. A new tool called DISSECT for analysing large genomic data sets using a Big Data approach. *Nat Commun* **6**(2015).
27. Stuart, A. & Ord, J.K. (eds.). *Kendall's advanced theory of statistics.*, 700 (Hoder Arnold, London, 1994).

Figure 1. Heritability estimates using SEM family-based models (self-reported data) and SNPs (self-reported data and medical records).

Black and grey sets show the three heritability estimates for each disease using SEM family-based models (self-reported data) and SNPs (self-reported data and medical records).

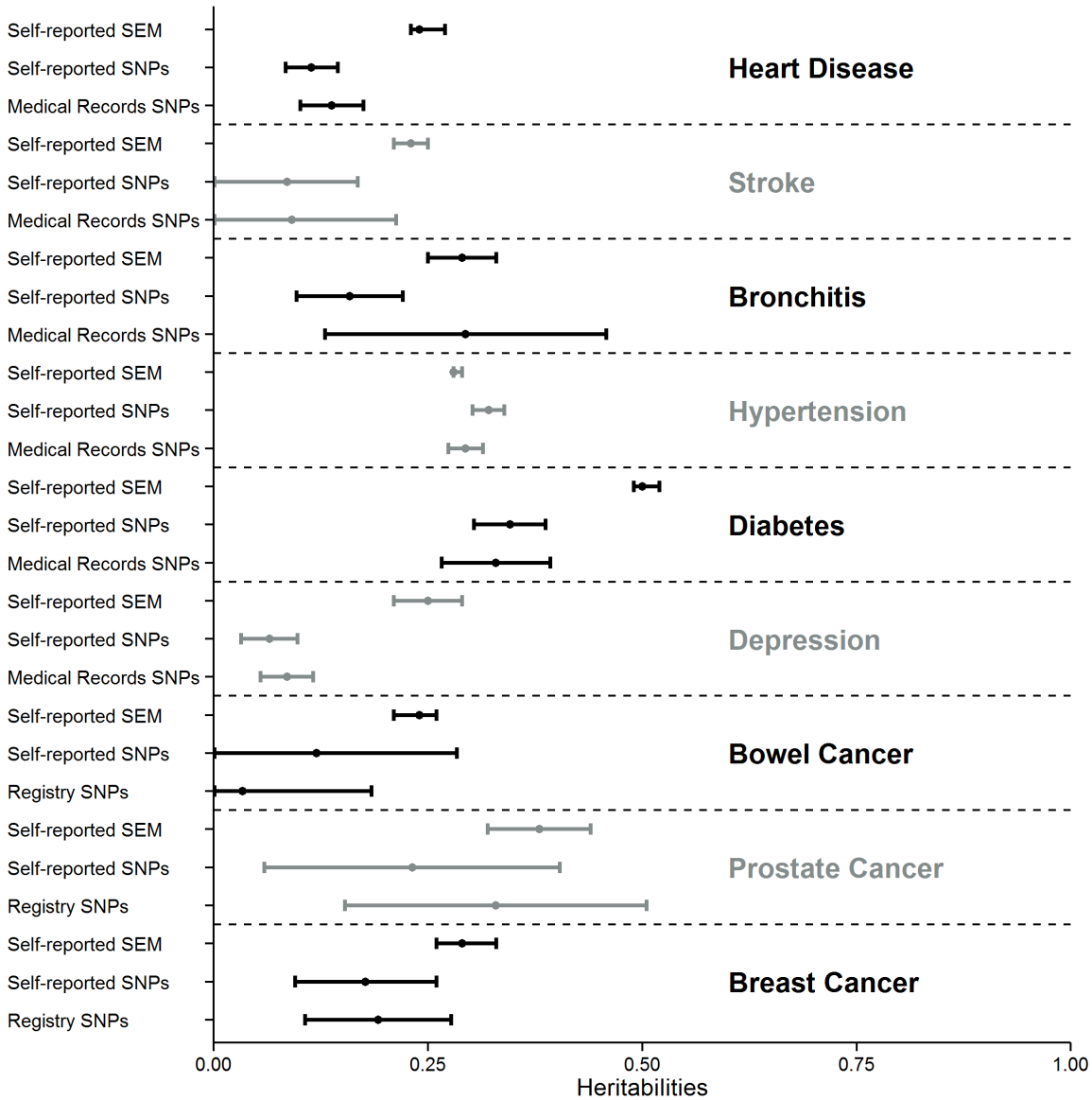


Table 1. Family-based heritability estimates not accounting for shared environmental effects calculated by Falconer's method and regression coefficients derived from different relative pairs.

Disease	h^2_{PO} (SE)	h^2_{SIB} (SE)	h^2_{PSIB} (SE)	b_{APO} (SE)	b_{ASIB} (SE)	b_{PAR} (SE)
Heart Disease	0.368 (0.005)	0.557 (0.018)	0.514 (0.010)	0.114 (0.026)	0.145 (0.108)	0.151 (0.003)
Stroke	0.162 (0.010)	0.305 (0.044)	0.260 (0.017)	-0.057 (0.054)	-	0.038 (0.004)
Bronchitis	0.420 (0.009)	0.501 (0.034)	0.567 (0.017)	0.169 (0.039)	0.338 (0.138)	0.108 (0.005)
Hypertension	0.366 (0.009)	0.691 (0.010)	0.477 (0.008)	0.035 (0.021)	0.190 (0.056)	0.203 (0.002)
Diabetes	0.474 (0.007)	0.692 (0.019)	0.485 (0.012)	0.067 (0.037)	0.185 (0.098)	0.109 (0.004)
Alzheimer's	0.238 (0.061)	-	0.349 (0.036)	-	-	0.060 (0.005)
Parkinson's	0.247 (0.038)	-	0.214 (0.053)	-	-	0.028 (0.013)
Depression	0.491 (0.007)	0.443 (0.019)	0.642 (0.013)	0.250 (0.036)	0.184 (0.083)	0.162 (0.005)
Lung cancer	0.117 (0.038)	-	0.314 (0.025)	-	-	0.119 (0.005)
Bowel cancer	0.260 (0.017)	0.387 (0.057)	0.300 (0.023)	0.171 (0.120)	-	0.032 (0.005)
Prostate cancer	0.361 (0.022)[¥]	0.707 (0.062)[¥]	0.321 (0.036)[¥]	-0.053 (0.183)	-	-
Breast cancer	0.287 (0.014)[Ψ]	0.393 (0.039)[Ψ]	0.301 (0.025)[Ψ]	0.144 (0.070)	-	-

h^2_{PO}: heritability estimates using data of parents and offspring; (SE): Standard errors between brackets; h^2_{SIB}: heritability estimates using data of siblings; h^2_{PSIB}: heritability estimates using data of parents and siblings of participants; b_{APO}: regression coefficient of parents on adopted offspring; b_{ASIB}: regression coefficient of adoptive siblings; b_{PAR} regression coefficient of parents of participants (partners); -: Effect was not estimated as there was less than one pair with both members affected; [¥]Only male-male pairs; [Ψ]Only female-female pairs.

Table 2. Genetic and environmental effects estimated using the parsimonious reduced SEM model.

Disease	Model	A (CI±0.95)	C (CI±0.95)	S (CI±0.95)	P (CI±0.95)	E (CI±0.95)
Heart Disease	ACSPE	0.27 (0.24-0.27)	0.08 (0.07-0.12)	0.08 (0.07-0.08)	0.06 (0.06-0.07)	0.51 (0.49-0.57)
Stroke	APE	0.23 (0.21-0.25)	-	-	0.04 (0.03-0.04)	0.73 (0.71-0.76)
Bronchitis	ACE	0.29 (0.25-0.33)	0.10 (0.10-0.11)	-	-	0.61 (0.60-0.64)
Hypertension	ACSPE	0.28 (0.28-0.29)	0.06 (0.06-0.06)	0.14 (0.14-0.14)	0.13 (0.12-0.13)	0.39 (0.38-0.39)
Diabetes	ASPE	0.50 (0.49-0.52)	-	0.11 (0.09-0.13)	0.07 (0.06-0.08)	0.32 (0.29-0.34)
Alzheimer's	ACE	0.25 (0.17-0.33)	0.05 (0.03-0.06)	-	-	0.70 (0.63-0.78)
Parkinson's	AE	0.26 (0.20-0.34)	-	-	-	0.74 (0.72-0.81)
Depression	ACE	0.25 (0.21-0.29)	0.15 (0.15-0.15)	-	-	0.60 (0.58-0.63)
Lung cancer	ACE	0.09 (0.02-0.14)	0.11 (0.09-0.13)	-	-	0.81 (0.75-0.86)
Bowel cancer	ACSE	0.24 (0.21-0.26)	0.03 (0.01-0.03)	0.06 (0.03-0.12)	-	0.67 (0.65-0.71)
Prostate cancer	ASE	0.38 (0.32-0.44)	-	0.19 (0.11-0.26)	-	0.43 (0.36-0.51)
Breast cancer	ASE	0.29 (0.26-0.33)	-	0.06 (0.01-0.10)	-	0.65 (0.60-0.69)

A: Additive genetic effects; C: Environmental effects common to the whole family; S: Sibling environmental effects; P: Partner environmental effects; E: Residual environmental effect; Confidence Interval at 95% between brackets. -: Parameter dropped from parsimonious reduced model.

Table 3. Heritability estimates of disease using common + rare SNPs and structural equation modelling (SEM) from self-reported data.

Disease	h^2_{C+R} ($CI_{95\%}$)	h^2_{SEM} ($CI_{95\%}$)	%(h^2_{C+R} /h^2_{SEM}) (SE)
Heart Disease	0.11 (0.08-0.15)	0.27 (0.24-0.27)	40.74 (9.79)
Stroke	0.09 (0.00-0.17)	0.23 (0.21-0.25)	39.13 (29.07)
Bronchitis	0.16 (0.10-0.22)	0.29 (0.25-0.33)	54.43 (15.72)
Hypertension	0.32 (0.30-0.34)	0.28 (0.28-0.29)	114.29 (3.29)
Diabetes	0.35 (0.30-0.39)	0.50 (0.49-0.52)	70.00 (5.23)
Depression	0.07 (0.03-0.10)	0.25 (0.23-0.27)	24.00 (13.79)
Bowel cancer	0.12 (0.00-0.28)	0.24 (0.21-0.26)	50.0 (49.75)
Prostate cancer	0.23 (0.06-0.40)	0.38 (0.32-0.44)	60.53 (30.42)
Breast cancer	0.18 (0.10-0.26)	0.29 (0.26-0.33)	62.07 (19.05)

h^2_{C+R}: Heritability estimates using SNPs in the liability scale; ($CI_{95\%}$): Confidence intervals; %(h^2_{C+R} /h^2_{SEM}): Percentage of SEM family-based estimate of heritability explained by SNPs. (SE): Standard error

ONLINE METHODS

UK Biobank Data

The UK Biobank database) includes 502,682 participants who were aged between 49-69 years when recruited between 2006 and 2010 from across UK to take part of the project. The study was approved by the National Research Ethics Committee (REC reference: 11/NW/0382). The participants filled several questionnaires about their lifestyle, environmental risk factors and medical history and gave their informed consent[4]. The comprehension and acceptability of each question, the time taken to complete each of them, and their response distributions were examined in pilot studies, which aided the final selection and presentation of suitable questions. Self-reported medical history was confirmed by a trained nurse and where necessary by a medical doctor. Moreover, a pre-visit questionnaire was provided to participants before attending the assessment center, this questionnaire afforded participants the opportunity to record personal information such as family history before the visit to the assessment center to minimize problems of recalling. These details were entered directly into the assessment center computer and the questionnaire was not retained[5]. The UK Biobank contains information on about 445 types of diseases and 81 cancers in participants and the familial medical history of twelve broadly defined diseases among blood and adoptive fathers, mothers and siblings. Participants were considered as adopted when they answer “Yes” to the question: “Were you adopted as a child?”.

Family pairs (parent-offspring, sib-sib, parent-sibs and partners) were characterized for these twelve diseases, which include different subcategories in participants. The diseases analyzed were: heart disease (twenty-five subcategories), stroke (three subcategories), chronic bronchitis (three subcategories), hypertension (two subcategories), diabetes (four subcategories), Alzheimer's disease, Parkinson's disease, severe depression, lung cancer (two subcategories), bowel cancer (five subcategories), prostate and breast cancers (Supplementary Table 1). Those participant who answered “not to know” or “prefer not to answer” when they were asked about the disease status of relatives were removed from the analyses. Disease

status of sibling was only considered when participants reported to have one sibling since they just had to report if at least one sibling had the corresponding disease and it was not possible to know how many siblings had suffered the disease when participants had more than one sibling. Disease status of 470,640 participants, 464,302 blood mothers, 459,716 blood fathers, 152,887 blood siblings, 4,962 adoptive mothers, 4,580 adoptive fathers and 1,819 adoptive siblings were used in the analyses. Those participants declared to have “white”, “British”, “Irish” or “Other white” ethnic background.

Medical Records

Data from medical records were used to test the accuracy of ten self-reported diseases. The type of medical record used to define a disease was different depending on the disease and was chosen because it was considered to be the best indicator of the disease available. Supplementary Table 2 shows the categories used to define each disease. There were available three kinds of medical records in the UK Biobank: data of hospitalization, medication/treatment and cancer register.

- Data of hospitalization. Summary of the distinct main diagnoses codes a participant has had recorded across all their episodes in hospital. Heart disease, stroke, bronchitis, diabetes and Parkinson’s and Alzheimer’s diseases were defined with records from this register.
- Medication/treatment. Medication self-reported by the participant used to treat the disease. Hypertension, diabetes and depression were defined with records form this registers.
- Cancer Register. Data from the UK Cancer Register was used to define the cancer diagnoses.

Accuracy of self-reported data and family health status

Accuracy of self-reported health status was evaluated estimating the sensitivity, specificity, positive (PPV), and negative (NPV) predictive values among self-reported data and medical records from cancer register, hospitalization records and medication. The sensitivity was estimated as the percentage of individuals who self-reported having a disease among all those

who appeared in the corresponding register as ill or taking the medication for the disease analyzed, the specificity was calculated as the percentage of those who self-reported being healthy for a particular disease among those who did not appear in the corresponding register or did not report taking medication for the corresponding disease. Positive predictive value (PPV) is the percentage of individuals who appeared in the corresponding register or were taking medication for a particular disease among those who self-reported having a disease, and the negative predictive value (NPV) is the proportion of those who did not appear in the registers or they did not report taking medication for the disease analyzed among those who did not report to have a particular disease. There were a total of 305,695 participants with hospitalization records that were used to estimate the accuracy of the self-reported phenotypes.

Prevalence

Prevalence of diseases in the UK Biobank were estimated as the number of people found to have a disease divided by the total number of individuals studied and their standard errors (SE) were estimated using the following formula:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

where *p* is the prevalence and n is the total number of individuals studied.

Relative Risks

Relative risks of disease in the UK Biobank were estimated as follow[21]:

$$RR = \frac{\frac{a}{a+b}}{\frac{a+c}{a+b+c+d}}$$

where *a* was the number of ill relatives of ill participants, *b* was the number of healthy relatives of ill participants, *c* was the number of ill relatives of healthy participants and *d* was the number of healthy relatives of healthy participants.

The relative risk of parents (RR_{PO}) and the relative risk of siblings (RR_{SIB}) were estimated using this formula. The relative risks of partners, who are parents of participant, (RR_{PAR}) was calculated in a similar way. The 95% confidence intervals (CI95%) were estimated as:

$$CI95\% = e^{log_e\,(RR)\pm 1.96s}$$

where RR is the corresponding relative risks and s is estimated as:

$$s = \sqrt{\frac{a^2+bc}{a(a+b)(a+c)} - \frac{1}{a+b+c+d}}$$

The minimum number of pairs in which both individuals are affected needed to estimate RR is one. In our dataset, the lowest number of pairs available to estimate RR was 33.

Heritability estimates

Diseases were treated as binary traits assumed to be determined by an underlying normal distribution of liability to disease. The correlation or regression among relatives (b) was used to estimate the heritability ($h^2 = 2b$) of liability to disease. Method 4 described by Falconer[1] was used to estimate b:

$$b = \frac{p_g\,(x_c - x_r)}{a_g}$$

where p_g is the prevalence of the disease in the relevant population within the UK Biobank, x_c is the deviation of the threshold of liability that defines disease status from the mean of relatives of healthy participants, x_r is the deviation of the threshold of liability that defines disease status from the mean of relatives of ill participants, and a_g is the mean liability deviation of the ill participants from the mean liability of the relevant population within the UK Biobank. The sampling variance (V_b) of b was estimated according to appendix C of

Falconer[1] and confirmed by bootstrapping. The minimum number of pairs in which both individuals are affected needed to estimate *b* was one. In our data set the lowest number of pairs available to estimate *b* was three (in the adoptive pairs).

Across generation differences in disease prevalence were taken into account using an appropriate control population for comparison. Since prevalences among genders were different, four estimates according gender pairs were estimated using this method, which allows controlling for differences in gender and age prevalence when the variance in mean liability is different. The following sets of relatives were used: parent-offspring, sib-sib and parent-sib of participants (blood and adoptive) except for prostate and breast cancers where we only estimated same gender correlations. Moreover, *b* was estimated among the parents of the participants. For each relationship class, the correlations or regressions obtained from the four gender-parings were combined into a single weighted mean (b_w), the weight being the reciprocal of the sampling variance of each regression coefficient. The sampling variance (V_{bw}) was calculated as the reciprocal of the sum of the weights and the standard error of the heritability was obtained as the square root of $4Vb_w$.

Genetic correlation

Genetic correlation (r_G) between sexes was calculated for all the diseases except for prostate cancer and breast cancer which are expressed mostly in one sex. The following formula was used[22]:

$$r_G = \sqrt{(b_{FEMALE\text{-}MALE} b_{MALE\text{-}FEMALE})/(b_{FEMALE\text{-}FEMALE} b_{MALE\text{-}MALE})}$$

where $b_{FEMALE\text{-}MALE}$ is the regression/correlation of mother-son or sister-brother $b_{MALE\text{-}FEMALE}$, the regression/correlation of father-daughter or brother-sister, $b_{FEMALE\text{-}FEMALE}$, the regression/correlation of mother-daughter or sister-sister and $b_{MALE\text{-}MALE}$ the regression/correlation of father-son or brother-brother.

Liability components

The liability to disease is the sum of genetic and different environmental effects. The distribution of the liability has a threshold value which differentiates between healthy and ill individuals. This threshold is based on the prevalence of the disease. As the prevalences are different in parents, siblings and participants, different thresholds must be assumed.

To estimate the liability parameters we can define the following structural equation:

$$L = A + C + S + P + E$$

where, A are genetic effects (assumed additive in the liability scale)[23]; C are environmental effects shared in common by all family members; S are environmental effects shared by siblings but not their parents which may include non-additive genetic effects; P are environmental effects shared among parents of participant (i.e. among partners) but not their children; and E are residual effects (including environmental effects specific to an individual and measurement error).

The correlations between each pair of blood and adoptive relatives for genetic and environmental components are set to fixed values according to their degree of genetic and environmental relationship. For example, blood parents-offspring pairs are correlated 0.5 for the genetic factors and 1 for common environmental effects. All the corresponding correlation values are shown in the Supplementary Figure 2. The relative importance of these components was evaluated using structural equation models (SEM) using OpenMx software version 1.4-3532[24].

Data of 210,787 blood and 4,184 adoptive families with one or two offspring (i.e. the participant and one sibling) were used to estimate liability components. A full model including all the effects (ACSPE) and all reduced models including genetic effects but removing one or more environmental effects were fitted. Each model was run 1,000 times and the run that converged with the maximum likelihood was chosen for model comparison. The relative fit of nested models was compared using hierarchic chi-square

tests because the difference between the likelihood for a reduced model and that for the full model is approximately distributed, as a chi-square with df = df(full model) - df(reduced model). For each disease we started with the simplest model and included more parameters until we obtained the most parsimonious but adequate model that did not fit the data significantly worse than the full model.

Simulations

We simulated pedigrees with the same structure of families as in the real data comprising 210,787 blood and 4,184 adoptive families. To simulate the diseases, the prevalences of each disease in fathers, mothers, participant and siblings were used together with the parameters obtained using the full model (Supplementary Table 8). The full model was fitted using OpenMx following the same procedure as with real data. Analyses with 10 simulation replicates for each disease were performed to estimate liability parameters. The means and standard deviations of the 10 replicates for each of the liability components were estimated. Model comparison for each replicate was carried out in the same way as with real data.

Genotype Quality Control

We use data from the genotyped individuals in phase 1 of the UK Biobank genotyping program. In this phase, 49,979 individuals were genotyped using the Affymetrix UK BiLEVE Axiom array and 102,750 individuals using the Affymetrix UK Biobank Axiom array. Further details regarding genotyping procedure and genotype calling protocols are available at the UK Biobank website. We excluded multi-allelic markers, SNPs with an overall missing rate higher than 2% or with a strong platform specific missing bias (Fisher's exact test, $P < 10^{-100}$). We also excluded individuals with a missing rate higher than 5%, with a self-reported sex different from the genetic sex estimated from the X chromosome inbreeding or those with an excess of heterozygosity according to the UK Biobank internal QC procedures.

A reduced dataset of 151,532 individuals remained after filtering. In addition to this, common and rare variants (i.e. with a MAF > 0.0036) and those that did not exhibit departure from Hardy-Weinberg equilibrium ($P < 10^{-50}$) in the unrelated (subset of 114,264 individuals with a relatedness below 0.0625) White-British cohort were kept. The genotype quality control and data filtering was performed using plink[25].

SNP heritability estimates

SNP heritability estimates were estimated in a subset of 114,264 individuals for nine out of the twelve diseases with a prevalence in the population higher than 0.50% (heart disease, stroke, chronic bronchitis, hypertension, diabetes, severe depression and bowel, prostate and breast cancers) using self-reported data and medical records.

To estimate the heritability for each disease and data set, the genetic relationship matrices (GRMs) were computed fitting simultaneously 525,242 SNPs in the following mixed lineal model:

$$y=X\beta+Wu+\epsilon$$

where y is the vector of phenotypes (diseases), β is the vector of fixed effects and covariates which included age of participant, the 20 first principal components and gender (except for prostate and breast cancer), u is the vector of SNP effects distributed as $u \sim N(0, I, \sigma^2_u)$, I is the identity matrix, and ϵ is a vector of residual effects distributed as $\epsilon \sim N(0, I\sigma^2_\epsilon)$. W is a genotype matrix defined by the equation:

$$W_{ik} = \frac{(s_{ik} - 2p_k)}{\sqrt{2p_k\,(1-p_k)}}$$

where s_{ik} is the number of copies of the reference allele for the SNP k of the individual i, and p_k is the frequency of the reference allele for the SNP k. Under this model, the variance of y is:

$$var(y) = A\sigma^2_g + I\sigma^2_\epsilon$$

Where A is the GRM, σ^2_g the genetic variance and σ^2_ε the residual variance. Variance components were estimated using restricted maximum likelihood (REML). These analyses were performed using DISSECT[26].

In addition to this, a two variance component model splitting the SNPs into 319,037 common SNPs (MAF>0.05) and 206,205 rare SNPs (0.0036<MAF<0.05) was fitted for each disease.

$$y=X\beta+W_{common}u_{common}+W_{rare}u_{rare}+\varepsilon$$

where, u_{common} and u_{rare} are the vectors of SNP effects for common and rare variants, respectively. W_{common} and W_{rare} are the genotype matrices defined for common and rare variants, respectively.
Under this model, the variance of y is:

$$var(y)=A_{common}\sigma^2_{g_{common}}+A_{rare}\sigma^2_{g_{rare}}+I\sigma^2_\varepsilon$$

where A_{common} and the A_{rare} are the GRMs computed using the common and rare variants, respectively. $\sigma^2_{gcommon}$ and σ^2_{grare} are the genetic variances explained by the common and rare variants, respectively.

The heritability estimates were transformed to the liability scale using the following equation:

$$h^2_L = h^2_{(0,1)} \frac{P(1-P)}{Z^2}$$

where h^2_L is the heritability in the liability scale is the heritability in the liability scale $h^2_{(0,1)}$ is the heritability in the observed scale obtained from the REML analyses, P is the prevalence of the disease in the cohort and Z is the height of the standard normal probability density function at the threshold that truncates the proportion P[23].
The percentage of SEM family-based estimates of heritability explained by SNPs was calculated as the ratio between h^2_{SNPs} and h^2_{SEM} multiplied by 100 and the standard error of the percentage was calculated according to Stuart *et al*[27] as:

$$SE(\%) = \sqrt{\left(\frac{h^2_{C+RSNPs}}{h^2_{SEM}}\right)^2 \left(\frac{\sigma^2_{h^2_{C+RSNPs}}}{h^4_{C+RSNPs}} + \frac{\sigma^2_{h^2_{SEM}}}{h^4_{SEM}}\right) - \frac{2COV(C+RSNPs,SEM)}{h^2_{C+RSNPs}\, h^2_{SEM}}} \times 100$$

where $h^2_{C+RSNPS}$ is the heritability explained by common and rare SNPs, h^2_{SEM} is the heritability using SEM family-based, $\sigma^2 h^2_{C+RSNPs}$ is the standard deviation of $h^2_{C+RSNPS}$, $\sigma^2 h^2_{SEM}$ is the standard deviation of $h^2_{SEM,}$ C+RSNPs are related to the distribution of the estimates of $h^2_{C+RSNPS}$ and SEM related to the distribution of the estimates of h^2_{SEM}. We cannot estimate *2COV(C+RSNPs,SEM)* and assume this value is equal to 0.

Testing of GWAS hits for self-reported and clinical definitions of disease

GWAS hits for breast cancer, prostate cancer, bowel cancer, Type 2 diabetes, Hypertension, Stroke and Cardiovascular Artery Disease were downloaded from the GWAS catalogue. In total, we found that 278 of these SNPs were genotyped in our array, and tested them for association with our self-reported and clinical definitions of disease (breast cancer, prostate cancer, bowel cancer, Type 2 diabetes, Hypertension, Stroke and Heart Disease) using a chi-square test as implemented in the plink2 option (--assoc)[25]. Significant SNPs at a p-value of 0.05 and 0.00018 (i.e. 0.05/278) were counted for the two definitions of disease (self-reported and clinical). Only the subset of genotyped samples with clinical information was used to compare the power of the two alternative phenotype definitions.

METHODS ONLY-REFERENCES

21. Risch, N. Linkage strategies for genetically complex traits. I. Multilocus models. American Journal of Human Genetics 46, 222-8 (1990).

22. Falconer, D.S. & Mackay, T.F. Introduction to Quantitative Genetics, (1996).

23. Dempster, E.R. & Lerner, I.M. Heritability of Threshold Characters. Genetics 35, 212-36 (1950).

24. Boker, S. et al. OpenMx: An Open Source Extended Structural Equation Modeling Framework. Psychometrika 76, 306-317 (2011).

25. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-75 (2007).

26. Canela-Xandri, O., Law, A., Gray, A., Woolliams, J.A. & Tenesa, A. A new tool called DISSECT for analysing large genomic data sets using a Big Data approach. Nat Commun 6(2015).

27. Stuart, A. & Ord, J.K. (eds.). Kendall's advanced theory of statistics., 700 (Hoder Arnold, London, 1994).

Supplementary note

Accuracy of self-reported data

The accuracy of self-reported health status was evaluated estimating the sensitivity, specificity, positive (PPV), and negative predictive values (NPV) among self-reported data and medical records from cancer register, hospitalization records and medication (Supplementary Tables 2 and 3). Specificity and NPV were 1 or close to 1 for all the diseases pointing out a negligible number of false negatives. Sensitivity and PPV was higher than 0.70 for all the cancer diseases and for hypertension and diabetes suggesting a high percentage of true positives. However, lower values of sensitivity and PPV were observed for heart disease, stroke, bronchitis and depression.

The cancer register is likely to be the most reliable database which likely captures the majority the cancers diagnosed in UK, however hospitalization records or medication may not contain all the cases, especially for less well recorded diseases. For instance, patients could be diagnosed with bronchitis but never have been hospitalized. In addition to this, a participant of the UK Biobank could report an episode of depression in the past and he/she could not report taking medication for this disease as would be already recovered from the disease when recruited.

Therefore the cancer register data is likely to be the most complete data set, whilst hospitalization and medication information is potentially incomplete. Given the accuracy of the self-reported data for cancer and the fact that medical history was confirmed by a trained nurse and where necessary by a medical doctor, it suggests that the self-reported data may too be of high quality.

Furthermore, self-reported data and medical records provided very similar estimates of SNP heritability for the nine diseases for which we estimated SNP heritabilities (Table 3 - Figure 1).

Supplementary Table 1. Different categories of disease in participants and their corresponding disease in relatives.

Relative disease	Participant disease
Heart disease	Heart Cardiac Problem
	Angina
	Heart attack/Myocardial infarction
	Heart failure/Pulmonary oedema
	Heart arrhythmia
	Atrial fibrillation
	Atrial flutter
	Wolf Parkinson White/wpw synd
	Irregular heart beat
	Sick sinus syndrome
	Svt/Supraventricular tachycardia
	Heart valve problem/heart murmur
	Mitral Valve disease
	Mitral Valve prolapse
	Mitral stenosis
	Mitral regurgitation
	Aortic valve disease
	Aortic stenosis
	Aortic regurgitation

	Cardiomyopathy
	Hypertrophic card
	Pericardial Problem
	Pericarditis
	Pericardial effusion
	Myocarditis
	Rheumatic Fever
Stroke	Stroke
	Subarachnoid Haemorrhage
	Ischaemic Haemorrhage
	Brain Haemorrhage
Chronic bronchitis/emphysema	Emphysema/Chronic Bronchitis
	Bronchitis
	Emphysema
	Alpha-1 Antitrypsin Deficiency
High blood pressure	Hypertension
	Essential Hypertension
	Gestational Hypertension/Preclampsia
Diabetes	Diabetes
	Gestational Diabetes
	Type 1 Diabetes

	Type 2 Diabetes Diabetes insipidus
Alzheimer's disease/dementia	Dementia/Alzheimer/Cognitive Impairment
Parkinson's disease	Parkinson’s disease
Severe depression	Severe depression
Lung cancer	Lung Cancer Small cell lung cancer Non-small cell lung cancer
Bowel cancer	Bowel cancer Small bowel cancer Large bowel cancer/colorectal cancer Colon Cancer/Sigmoid cancer Appendix cancer Rectal cancer
Prostate cancer	Prostate cancer
Breast cancer	Breast cancer

Supplementary Table 2. Medical records used to evaluate accuracy of self-reported data.

Self –reported disease	Medical Record	Number
Heart disease	Cardiac arrhythmias	9,676
	Myocardial Infarction	6,573
	Angina pectoris	5,452
	Heart failure	2,943
	Non-rheumatic valve disorders	2,855
	Cardiomyopathy	655
	Acute pericarditis	549
	Other ill-defined heart diseases	429
	Pulmonary oedema	328
	Heart disease, unspecified	189
	Myocarditis	27
	Acute rheumatic fever	14
Stroke	Cerebral infarction	1,855
	Stroke, not specified as haemorrhage or infarction	918
	Sequelae of cerebrovascular disease	828
	Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction	667
	Subarachnoid haemorrhage	656

	Intracerebral haemorrhage	433
	Cerebrovascular disease, unspecified	579
	Other non-traumatic intracranial haemorrhage	223
	Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction	30
Chronic bronchitis/emphysema	Simple and mucopurulent chronic bronchitis	899
	Bronchitis, not specified as acute or chronic	421
	Emphysema	186
	Unspecified chronic bronchitis	14
High blood pressure	Ramipril	21,885
	Amlodipine	19,074
	Atenolol	18,378
	Bisoprolol	6,915
	Perindopril	6,748
	Candesartan cilexetil	5,027
	Losartan	4,271
	Felodipine	3,961
	Enalapril	3,597
	Irbesartan	3,345
	Propranolol	3,137
	Lercanidipine	1,137
	Diltiazem	954

Metoprolol	898
Verapamil	789
Telmisartan	738
Sotalol	728
Olmesartan	549
Lacidipine	490
Losartan Potassium+Hydrochlorothiazide 50mg/12.5mg Tablet	373
Bisoprolol Fumarate+Hydrochlorothiazide 10mg/6.25mg Tablet	299
Eprosartan	273
Trandolapril	271
Captopril	206
Quinapril	204
Irbesartan+Hydrochlorothiazide 150mg/12.5mg Tablet	109
Valsartan+Hydrochlorothiazide 80mg/12.5mg Tablet	71
Atenolol+bendroflumethiazide	59
Nicardipine	57
Propranolol Hydrochloride+Bendrofluazide 80mg/2.5mg Capsule	51
Sotalol Hydrochloride+Hydrochlorothiazide 80mg/12.5mg Tablet	38
Atenolol+bendrofluazide	27

	Cilazapril	22
	Metoprolol Tartrate+Chlorthalidone 100mg/12.5mg Tablet	22
	Acebutolol	18
	Oxprenolol	16
	Pindolol	12
	Atenolol+chlorthalidone	10
	Atenolol+chlortalidone	8
	Atenolol+Nifedipine 50mg/20mg M/R Capsule	5
	Timolol Maleate+Bendrofluazide 10mg/2.5mg Tablet	5
	Metoprolol Tartrate+Hydrochlorothiazide 100mg/12.5mg Tablet	4
	Captopril+Hydrochlorothiazide 25mg/12.5mg Tablet	2
	Atenolol+co-amilozide	3
	Timolol Maleate+Co-Amilozide 10mg/2.5mg/25mg Tablet	3
Diabetes	Type 2 diabetes	13,408
	Other specified diabetes mellitus	2,630
	Insulin product	563
	Type 1 diabetes	366
	Pioglitazone	206
	Rosiglitazone	140

	Glucophage 500mg tablet	110
	Unspecified diabetes mellitus	107
	Glimepiride	68
	Actos 15mg tablet	15
	Repaglinide	13
	Nateglinide	12
	Avandia 4mg tablet	11
	Amaryl 1mg tablet	5
	Glipizide product	1
Alzheimer's Disease	Alzheimer's Dementia	90
Parkinson's Disease	Parkinson's Disease	569
Severe depression	Amitriptyline	8,968
	Citalopram	8,063
	Fluoxetine	5,502
	Sertraline	1,986
	Venlafaxine	1,917
	Paroxetine	1,591
	Prozac 20mg Capsule	644
	Seroxat 20mg Tablet	627
	Duloxetine	403
	Clomipramine	315

	Nortriptyline	277
	Lofepramine	273
	Efexor 37.5mg Tablet	242
	Imipramine	240
	Cipramil 10mg Tablet	162
	Amitriptyline Hydrochloride+Perphenazine 10mg/2mg Tablet	110
	Lustral 50mg Tablet	66
	Cymbalta 30mg Gastro-Resistant Capsule	43
	Anafranil 10mg Capsule	39
	Phenelzine	31
	Moclobemide	28
	Yentreve 20mg Gastro-Resistant Capsule	28
	Allegron 10mg Tablet	6
	Nardil 15mg Tablet	4
	Gamanil 70mg Tablet	2
	Manerix 150mg Tablet	2
	Tofranil 10mg Tablet	1
	Amitriptyline+Chlordiazepoxide 12.5mg/5mg Capsule	1
	Tryptizol 10mg Tablet	1
Lung cancer	Malignant neoplasm of bronchus and lung	281
	Neoplasma Trachea, bronchus and lung	31

	Neoplasm of pleura	2
	Secondary malignant neoplasm of lung	1
Bowel cancer	Malignant neoplasm of colon	1,390
	Malignant neoplasm of rectum	1,072
	Neoplasm of uncertain behaviour	186
	Malignant neoplasm of small intestine	55
	Carcinoma in situ of rectum	38
	Carcinoma in situ of colon	35
	Benign neoplasm of colon, rectum, anus and anal canal	5
	Carcinoma in situ of rectosigmoid junction	3
Prostate cancer	Carcinoma in situ of prostate	3,126
	Malignant neoplasm of prostate	169
	Neoplasm of uncertain behavior of prostate	9
Breast cancer	Malignant neoplasm of breast	7,589
	Carcinoma in situ of breast	1,197
	Neoplasm of uncertain or unknown behaviour breast	29

Supplementary Table 3. Sensitivity, specificity, positive (PPV) and negative (NPV) predictive values among self-reported data and medical record used.

Disease	Medical Records	Sensitivity ($CI_{95\%}$)	Specificity ($CI_{95\%}$)	PPV ($CI_{95\%}$)	NPV ($CI_{95\%}$)
Heart Disease	Hospitalization	0.65 (0.64-0.65)	0.97 (0.97-0.97)	0.72 (0.71-0.72)	0.96 (0.96-0.96)
Stroke	Hospitalization	0.55 (0.54-0.57)	0.99 (0.99-0.99)	0.42 (0.41-0.44)	0.99 (0.99-0.99)
Bronchitis	Hospitalization	0.43 (0.41-0.44)	0.98 (0.98-0.98)	0.15 (0.14-0.16)	1.00 (1.00-1.00)
Hypertension	Medication	0.86 (0.85-0.87)	0.90 (0.90-0.90)	0.71 (0.71-0.71)	0.96 (0.96-0.96)
Diabetes	Hospitalization & Medication	0.85 (0.84-0.86)	0.98 (0.98-0.98)	0.72 (0.71-0.73)	0.99 (0.99-0.99)
Alzheimer's	Hospitalization	0.14 (0.09-0.18)	1.00 (1.00-1.00)	0.22 (0.15-0.30)	1.00 (1.00-1.00)
Parkinson's	Hospitalization	0.75 (0.72-0.79)	1.00 (1.00-1.00)	0.67 (0.63-0.70)	1.00 (1.00-1.00)
Depression	Medication	0.52 (0.51-0.52)	0.98 (0.98-0.98)	0.63 (0.62-0.64)	0.97 (0.96-0.97)
Lung cancer	Cancer Register	0.83 (0.79-0.87)	1.00 (1.00-1.00)	0.70 (0.65-0.74)	1.00 (1.00-1.00)
Bowel cancer	Cancer Register	0.83 (0.82-0.85)	1.00 (1.00-1.00)	0.85 (0.83-0.86)	1.00 (1.00-1.00)

Prostate cancer	Cancer Register	0.91 (0.90-0.92)	1.00 (1.00-1.00)	0.91 (0.90-0.92)	1.00 (1.00-1.00)
Breast cancer	Cancer Register	0.97 (0.97-0.97)	1.00 (1.00-1.00)	0.94 (0.94-0.95)	1.00 (1.00-1.00)

$CI_{95\%}$: Confidence intervals.

Supplementary Table 4. Prevalence percentages and standard errors of the analysed diseases in the UK Biobank.

	Prevalence % (SE)								
Disease	**Participants**	**Men**	**Women**	**Parents**	**Fathers**	**Mothers**	**Siblings**	**Brothers**	**Siste**
Heart Disease	7.19 (0.04)	10.09 (0.06)	4.77 (0.04)	23.91 (0.04)	29.09 (0.07)	18.72 (0.06)	4.86 (0.05)	6.99 (0.09)	2.56 (0.0
Stroke	1.45 (0.02)	1.86 (0.03)	1.10 (0.02)	13.51 (0.04)	13.72 (0.05)	13.30 (0.05)	1.42 (0.03)	1.09 (0.04)	1.73 (0.0
Bronchitis	2.16 (0.02)	2.34 (0.03)	2.01 (0.03)	7.86 (0.03)	10.17 (0.04)	5.54 (0.03)	1.30 (0.03)	1.40 (0.04)	1.20 (0.0
Hypertension	26.63 (0.06)	30.29 (0.10)	23.57 (0.08)	24.17 (0.04)	19.85 (0.06)	28.50 (0.07)	13.94 (0.09)	13.47 (0.13)	14.37 (0.1
Diabetes	4.71 (0.03)	6.48 (0.05)	3.23 (0.03)	8.57 (0.03)	8.41 (0.04)	8.72 (0.04)	4.21 (0.05)	5.05 (0.08)	3.30 (0.0
Alzheimer's	0.02 (0.002)	0.03 (0.01)	0.02 (0.01)	6.10 (0.02)	4.21 (0.03)	7.99 (0.04)	0.21 (0.01)	0.19 (0.02)	0.24 (0.0

Parkinson’s	0.17 (0.01)	0.24 (0.01)	0.12 (0.01)	1.87 (0.01)	2.21 (0.02)	1.53 (0.02)	0.24 (0.01)	0.32 (0.02)	0.16 (0.0
Depression	5.71 (0.03)	4.28 (0.04)	6.90 (0.05)	4.77 (0.02)	3.35 (0.03)	6.19 (0.04)	4.17 (0.05)	4.59 (0.08)	3.78 (0.0
Lung cancer	0.09 (0.01)	0.11 (0.01)	0.07 (0.01)	6.02 (0.02)	8.19 (0.04)	3.86 (0.03)	0.77 (0.02)	0.92 (0.03)	0.61 (0.0
Bowel cancer	0.58 (0.01)	0.71 (0.02)	0.46 (0.01)	5.05 (0.02)	5.24 (0.03)	4.85 (0.03)	0.99 (0.03)	1.08 (0.04)	0.89 (0.0
Prostate cancer	-	1.54 (0.03)	-	-	6.79 (0.05)	-	-	1.49 (0.04)	
Breast cancer	2.31 (0.02)	0.02 (0.01)	4.22 (0.04)	-	-	7.85 (0.04)	2.17 (0.04)	0.04 (0.01)	4.47 (0.0

SE: Standard errors; -: There were not data for the corresponding disease.

Supplementary Table 5. Genetic correlations among genders.

Illness	r_{GPO} (SE)	r_{GSIB} (SE)
Heart Disease	0.81 (0.01)	0.90 (0.01)
Stroke	0.78 (0.02)	1.24 (0.57)
Bronchitis	0.87 (0.01)	0.80 (0.05)
Hypertension	0.78 (0.01)	0.95 (0.01)
Diabetes	0.93 (0.01)	0.92 (0.01)
Alzheimer's	0.83 (0.07)	-
Parkinson's	1.03 (0.01)	-
Depression	0.96 (0.01)	0.93 (0.02)
Lung cancer	2.40 (1.34)	-
Bowel cancer	0.73 (0.02)	0.99 (0.01)

r_{GPO}: Genetic correlations using parent-offspring pairs;

r_{GSIB}: Genetic correlation using sibling pairs. SE: Standard errors,

-: There was no pair with the two members affected to estimate the genetic correlation

Supplementary Table 6. Heritability estimates from published twin studies.

Disease	h^2 ($CI_{95\%}$)	Study[§]
Heart Disease	0.39 (0.29-0.49) ♂	Zdravkovic *et al.* (2007)
	0.43 (0.08-0.51) ♀	
Stroke	0.41 (0.24-0.56)	Korja *et al.* (2010)
Bronchitis	0.40 (0.33–0.47) ♂	Hallberg *et al.* (2007)
	0.38 (0.12–0.47) ♀	
Hypertension	0.39 (0.2 – 0.5) DBP	Iliadou *et al.* (2002)
	0.52 (0.37 – 0.65) SBP	
Diabetes	0.64[¥]	Kaprio *et al.* (1992)
Alzheimer's	0.79 (0.67-0.88)	Gatz *et al.* (2006)
Parkinson's	0.36 (0.28-0.44) ♂	Polderman *et al.* (2015)
Depression	0.37 (0.33-0.42)	Sullivan *et al.* (2010)
Lung cancer	0.26 (0.00–0.49)	Lichtenstein *et al.* (2010)
Bowel cancer	0.35 (0.10–0.48)	Lichtenstein *et al.* (2010)
Prostate cancer	0.42 (0.29–0.50)	Lichtenstein *et al.* (2010)
Breast cancer	0.27 (0.04–0.41)	Lichtenstein *et al.* (2010)

h^2: Heritability estimates; ($CI_{95\%}$): Confidence intervals; [§] Studies cited in main text and bibliography; ♂: Males; ♀: Females [¥] Confidence intervals not shown in the paper.

Supplementary Table 7. Genetic and environmental effects estimated using the full and parsimonious reduced SEM models.

Disease	Model	LL	*p*-value	A	C	S	P	E
Heart Disease	ACSPE	1669870.83[¥]	<0.05[*]	0.27	0.08	0.08	0.06	0.51
Stroke	ACSPE	1048621.32	0.16	0.23	0.00	0.05	0.04	0.69
	APE	1048624.91		0.23	-	-	0.04	0.73
Bronchitis	ACSPE	739568.10	0.24	0.29	0.10	0.03	0.00	0.59
	ACE	739570.99		0.29	0.10	-	-	0.61
Hypertension	ACSPE	2278607.50[¥]	<0.05[*]	0.28	0.06	0.14	0.13	0.39
Diabetes	ACSPE	1002661.33	0.60	0.47	0.02	0.11	0.05	0.35
	ASPE	1002661.60		0.50	-	0.11	0.07	0.32
Alzheimer's	ACSPE	558921.77	1	0.28	0.03	0.03	0.01	0.65
	ACE	558921.21		0.25	0.05	-	-	0.70
Parkinson's	ACSPE	252436.39	0.37	0.20	0.03	0.05	0.00	0.72
	AE	252439.54		0.26	-	-	-	0.74
Depression	ACSPE	839646.58	1	0.25	0.15	0.00	0.00	0.60
	ACE	839646.58		0.25	0.15	-	-	0.60
Lung cancer	ACSPE	522617.83	0.71	0.21	0.05	0.05	0.06	0.64
	ACE	522618.49		0.09	0.11	-	-	0.81
Bowel cancer	ACSPE	544144.54	1	0.23	0.03	0.06	0.00	0.68
	ACSE	544144.54		0.24	0.03	0.06	-	0.67
Prostate cancer	ACSE	98725.19	1	0.38	0.00	0.19	-	0.43
	ASE	98725.19		0.38	-	0.19	-	0.43
Breast cancer	ACSE	204255.53	1	0.29	0.00	0.06	-	0.65
	ASE	204255.53		0.29	-	0.06	-	0.65

LL: -2log likelihood; [¥] for heart disease and hypertension only the full model is shown because it was the most parsimonious; *p*-value corresponding to the difference between LL, for heart disease and hypertension the *p-value* was always lower than 0.05 for all the comparisons; [*]all the reduced models were significantly worse than the full one; A: Additive genetic effects; C: Environmental effects common to the whole family; S: Sibling environmental effects; P: Partner environmental effects; E: Residual environmental effect;

-: effect not included in the model

Supplementary Table 8. Parameters used in the simulations of disease.

Disease	Model	Liability components					Prevalences			
		A	C	S	P	E	Father	Mother	Participant	Sibling
Heart Disease	ACSPE	0.27	0.08	0.08	0.06	0.51	0.291	0.187	0.0719	0.049
Stroke	ASPE	0.23	0.00	0.05	0.04	0.69	0.137	0.133	0.0145	0.014
Bronchitis	ACSE	0.29	0.10	0.03	0.00	0.59	0.102	0.055	0.0216	0.013
Hypertension	ACSPE	0.28	0.06	0.14	0.13	0.39	0.199	0.285	0.2663	0.139
Diabetes	ACSPE	0.47	0.02	0.11	0.05	0.35	0.084	0.087	0.0471	0.042
Alzheimer’s	ACSPE	0.28	0.03	0.03	0.01	0.65	0.042	0.080	0.0002	0.002
Parkinson’s	ACSE	0.20	0.03	0.05	0.00	0.72	0.022	0.015	0.0017	0.002
Depression	ACE	0.25	0.15	0.00	0.00	0.60	0.034	0.062	0.0570	0.042
Lung cancer	ACSPE	0.21	0.05	0.05	0.06	0.64	0.082	0.039	0.0009	0.008
Bowel cancer	ACSE	0.23	0.03	0.06	0.00	0.68	0.052	0.049	0.0058	0.010
Prostate cancer	ASE	0.38	0.00	0.19	-	0.43	0.068	-	0.0154	0.015

Breast cancer	ASE	0.29	0.00	0.06	-	0.65	-	0.079	0.0422	0.045

A: Additive genetic effects; C: Environmental effects common to the whole family; S: Sibling environmental effects; P: Partner environmental effects; E: Residual environmental effect; -: effect not included in the model or prevalence not estimated because the disease is limited to one sex

Supplementary Table 10. Number of simulation replicates where the SEM recovers the model used to simulate data, and mean and standard deviation of the parameter estimates.

Diseases	Model	Replicates	A	C	S	P	E
Heart Disease	ACSPE	5	0.29 (0.02)	0.08 (0.01)	0.08 (0.01)	0.06 (0.01)	0.48 (0.02)
Stroke	ASPE	4	0.26 (0.01)	-	0.06 (0.02)	0.04 (0.01)	0.65 (0.01)
Bronchitis	ACSE	1	0.33 (ne)	0.08 (ne)	0.04 (ne)	-	0.55 (ne)
Hypertension	ACSPE	8	0.32 (0.04)	0.07 (0.02)	0.16 (0.00)	0.15 (0.02)	0.30 (0.03)
Diabetes	ACSPE	0	-	-	-	-	-
Alzheimer's	ACSPE	0	-	-	-	-	-
Parkinson's	ACSE	0	-	-	-	-	-
Depression	ACE	10	0.25 (0.02)	0.15 (0.01)	-	-	0.60 (0.01)
Lung cancer	ACSPE	0	-	-	-	-	-
Bowel cancer	ACSE	2	0.25 (0.01)	0.03 (0.01)	0.08 (0.01)	-	0.64 (0.01)
Prostate cancer	ASE	10	0.43 (0.03)		0.16 (0.02)	-	0.41 (0.03)

Breast cancer	ASE	3	0.30 (0.02)	-	0.08 (0.02)	-	0.62 (0.04)	

SD: Between brackets; ne: non-estimable; - not recovered parameter

Supplementary Table 11. Number of simulation replicates where the SEM does not recover the model used to simulate data, and the mean and standard deviation of the parameter estimates.

Diseases	Replicates	A	C	S	P	E
Heart Disease	5	0.28 (0.16)	0.09 (0.08)	0.09 (0.00)	0.06 (0.08)	0.48 (0.16)
Stroke	6	0.25 (0.01)	0 (0.00)	0 (0.00)	0.04 (0.01)	0.71 (0.02)
Bronchitis	9	0.41 (0.11)	0.05 (0.05)	0 (0.00)	0.06 (0.05)	0.49 (0.11)
Hypertension	2	0.46 (0.00)	0 (0.00)	0.16 (0.00)	0.22 (0.01)	0.17 (0.01)
Diabetes	10	0.53 (0.05)	0.01 (0.03)	0.11 (0.01)	0.08 (0.03)	0.28 (0.05)
Alzheimer’s	10	0.32 (0.07)	0.02 (0.02)	0 (0.00)	0.03 (0.02)	0.64 (0.07)

Parkinson's	10	0.27 (0.02)	0 (0.00)	0 (0.00)	0.01 (0.03)	0.73 (0.04)
Depression	0	-	-	-	-	-
Lung cancer	10	0.07 (0.03)	0.12 (0.01)	0.08 (0.09)	-	0.72 (0.08)
Bowel cancer	8	0.30 (0.01)	0.00 (0.01)	0.02 (0.04)	0.02 (0.02)	0.66 (0.05)
Prostate cancer	0	-	-	-	-	-
Breast cancer	7	0.30 (0.02)	0 (0.00)	0 (0.00)	-	0.70 (0.02)

SD: Between brackets; - not recovered parameter

Supplementary Table 9. Mean of parameter estimates from SEM under the full model of the ten replicates for each disease simulated.

Diseases	A	C	S	P	E
Heart Disease	0.28 (0.06)	0.09 (0.03)	0.08 (0.01)	0.06 (0.03)	0.48 (0.05)
Stroke	0.24 (0.03)	0.00 (0.01)	0.05 (0.02)	0.04 (0.02)	0.67 (0.04)
Bronchitis	0.37 (0.08)	0.06 (0.04)	0.03 (0.01)	0.03 (0.04)	0.51 (0.08)
Hypertension	0.34 (0.05)	0.06 (0.02)	0.16 (0.01)	0.16 (0.02)	0.28 (0.05)
Diabetes	0.53 (0.03)	0.01 (0.02)	0.11 (0.01)	0.07 (0.02)	0.28 (0.04)
Alzheimer's	0.33 (0.05)	0.01 (0.02)	0.06 (0.07)	0.03 (0.02)	0.57 (0.10)
Parkinson's	0.25 (0.02)	0.00 (0.01)	0.03 (0.04)	0.04 (0.03)	0.68 (0.05)
Depression	0.28 (0.04)	0.13 (0.02)	0.01 (0.01)	0.02 (0.02)	0.56 (0.04)
Lung cancer	0.12 (0.06)	0.09 (0.03)	0.12 (0.07)	0.03 (0.03)	0.64 (0.09)
Bowel cancer	0.29 (0.03)	0.01 (0.01)	0.06 (0.06)	0.03 (0.03)	0.62 (0.04)
Prostate cancer	0.25 (0.19)	0.09 (0.09)	0.15 (0.03)	-	0.51 (0.09)
Breast cancer	0.23 (0.06)	0.03 (0.03)	0.05 (0.03)	-	0.69 (0.04)

SD: Between brackets, -: effect not included in the model or prevalence not estimated because the disease is limited to one sex

Supplementary Table 12. Mean of parameter estimates from SEM of the AE model of the ten replicates for each disease simulated under the full model.

Diseases	A	E
Heart Disease	0.48 (0.01)	0.52 (0.01)
Stroke	0.25 (0.01)	0.75 (0.01)
Bronchitis	0.50 (0.01)	0.50 (0.01)
Hypertension	0.51 (0.01)	0.49 (0.01)
Diabetes	0.58 (0.01)	0.42 (0.01)
Alzheimer's	0.36 (0.03)	0.64 (0.03)
Parkinson's	0.26 (0.02)	0.74 (0.02)
Depression	0.55 (0.01)	0.45 (0.01)
Lung cancer	0.31 (0.03)	0.69 (0.03)
Bowel cancer	0.31 (0.02)	0.69 (0.02)
Prostate cancer	0.48 (0.03)	0.52 (0.03)
Breast cancer	0.31 (0.02)	0.67 (0.10)

SD: Between brackets

Supplementary Table 13. . Genetic and environmental effects estimated using the AE model in the UK Biobank data.

Disease	A	E
Heart Disease	0.44	0.56
Stroke	0.23	0.77
Chronic Bronchitis	0.48	0.52
Hypertension	0.45	0.55
Diabetes	0.54	0.46
Alzheimer's	0.35	0.65
Parkinson's	0.26	0.74
Depression	0.55	0.45
Lung cancer	0.30	0.70
Bowel cancer	0.30	0.70
Prostate cancer	0.43	0.56
Breast cancer	0.31	0.69

Supplementary Table 14. Heritability estimates of the self-reported data when modelling all SNPs in a single variance component (C+R), or splitting common SNPs and rare SNP to estimate two variance components simultaneously.

Disease	h^2_{C+R} $(CI_{95\%})_L$	h^2_C $(CI_{95\%})_L$	h^2_R $(CI_{95\%})_L$	$\%h^2_{C+R}/h^2_{SEM}$ (SE)	$\%(h^2_{C+} h^2_R)/h^2_{SEM}$ (SE)
Heart Disease	0.114 (0.084-0.145) ¥	0.088 (0.064-0.111)¥	0.007 (0.000-0.035)	40.74 (9.79)	35.21 (77.02)
Stroke	0.086 (0.004-0.168) ¥	0.043 (0.000-0.095)	0.026 (0.000-0.092)	39.13 (29.07)	30.07 (37.20)
Bronchitis	0.159 (0.097-0.221) ¥	0.149 (0.091-0.208) ¥	0.011 (0.000-0.077)	54.43 (15.72)	55.48 (25.12)
Hypertension	0.321 (0.302-0.339) ¥	0.261 (0.245-0.277)¥	0.026 (0.010-0.042)¥	114.29 (3.29)	102.50 (6.26)
Diabetes	0.346 (0.304-0.387) ¥	0.277 (0.243-0.312) ¥	0.032 (0.000-0.069)	70.00 (5.23)	61.86 (77.33)
Depression	0.065 (0.032-0.098) ¥	0.062 (0.035-0.088) ¥	0.000 (0.000-0.031)	24.00 (13.79)	24.62 (17.98)
Bowel cancer	0.120 (0.000-0.284)	0.067 (0.000-0.154)	0.000 (0.000-0.107)	50.0 (49.75)	27.98 (60.55)
Prostate cancer	0.232 (0.059-0.404) ¥	0.140 (0.027-0.253) ¥	0.030 (0.000-0.169)	60.53 (30.42)	44.61 (41.12)
Breast cancer	0.177 (0.095-0.260) ¥	0.112 (0.045-0.178) ¥	0.066 (0.000-0.145)	62.07 (19.05)	61.20 (23.59)

h^2_{C+R}: Common+rare variants; $h^2_{C:}$ Common variants, $h^2_{R:}$ Rare variants; $L_:$ Liability scale; $h^2_{SEM:}$ SEM heritability estimates; $CI_{95\%:}$ Confidence intervals at 95%; SE: Standard errors, ¥; *p-value* < 0.05 to test the heritability is significantly different from zero.

Supplementary Table 15. Heritability estimates from medical records data when modelling all SNPs in a single variance component (C+R), or splitting common SNPs and rare SNP to estimate two variance components simultaneously.

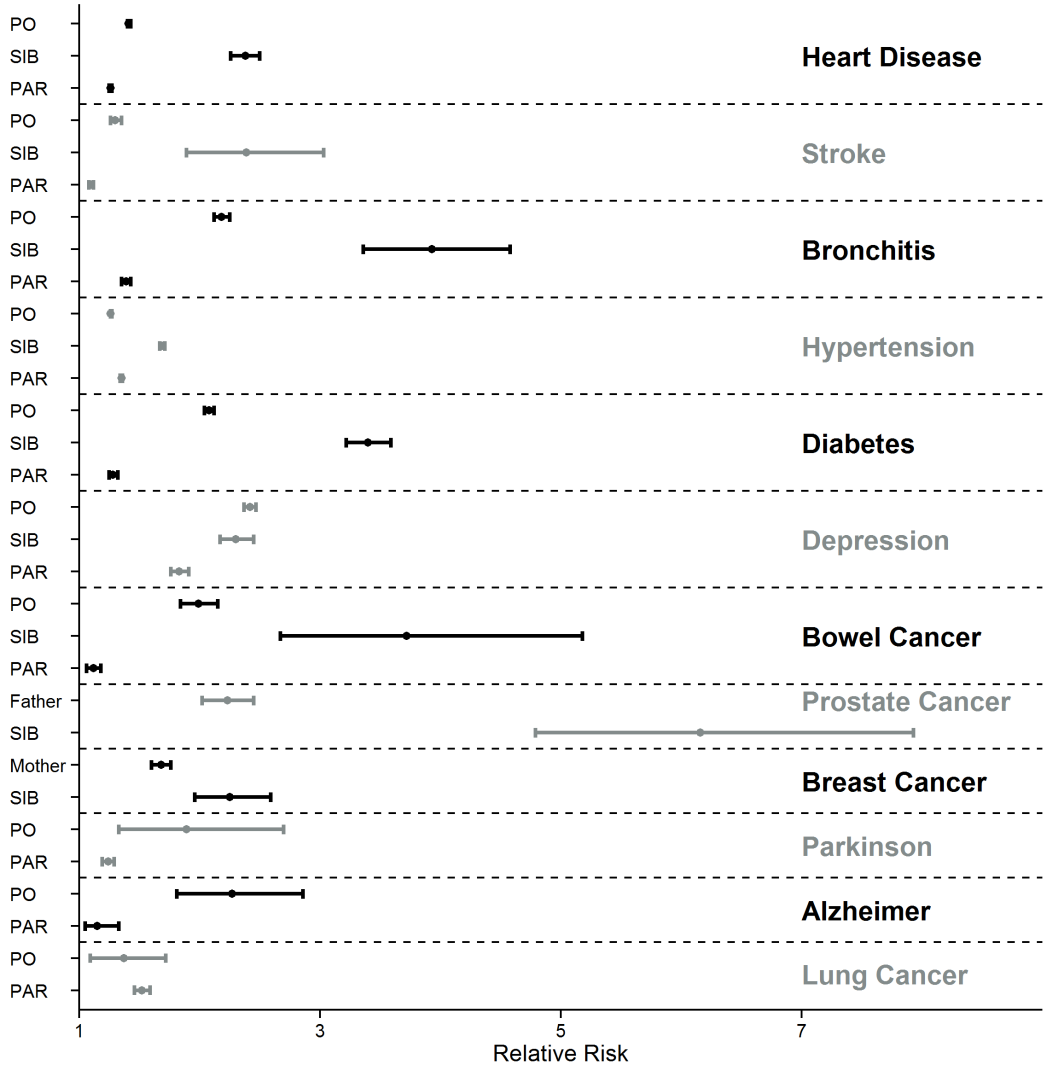
Disease	h^2_{C+R} $(CI_{95\%})_L$	h^2_C $(CI_{95\%})_L$	h^2_R $(CI_{95\%})_L$
Heart Disease	0.138 (0.101-0.175) ¥	0.110 (0.080-0.140) ¥	0.012 (0.000-0.046)
Stroke	0.091 (0.000-0.213)	0.105 (0.000-0.234)	0.000 (0.000-0.165)
Bronchitis	0.294 (0.130-0.458) ¥	0.259 (0.118-0.399) ¥	0.008 (0.000-0.170)
Hypertension	0.294 (0.274-0.314) ¥	0.231 (0.214-0.247) ¥	0.034 (0.017-0.050) ¥
Diabetes	0.329 (0.266-0.393) ¥	0.265 (0.216-0.314) ¥	0.000 (0.000-0.055)
Depression	0.086 (0.055-0.116) ¥	0.061 (0.036-0.086) ¥	0.021 (0.000-0.050)
Bowel cancer	0.034 (0.000-0.184)	0.000 (0.000-0.076)	0.055 (0.000-0.161)
Prostate cancer	0.329 (0.153-0.505) ¥	0.199 (0.082-0.316) ¥	0.049 (0.000-0.188)
Breast cancer	0.192 (0.107-0.277) ¥	0.110 (0.044-0.176) ¥	0.083 (0.000-0.162) ¥

$h^2_{C+R:}$ Common+rare variants; $h^2_{C:}$ Common variants, $h^2_{R:}$ Rare variants; $L_:$ Liability scale; ¥; *p-value* < 0.05 to test the heritability is significantly different from zero.

Supplementary Table 16. Number of significant GWAS hits using self-reported and clinical definitions of disease.

		Cases	Controls	Total	Number GWAS hits tested	Number significant at α = 0.05	Number significant at α = 0.00018	Overlap at α = 0.05	Overlap at α = 0.00018
Breast cancer	Self-reported	2425	57592	60017	48	27	11		
	Cancer registry	2339	57678	60017	48	22	10	22	10
Prostate cancer	Self-reported	833	53410	54243	42	21	13		
	Cancer registry	835	53408	54243	42	23	12	21	10
Colorectal cancer	Self-reported	675	113585	114260	23	8	1		
	Clinical/Registry	687	113573	114260	23	8	1	7	1
Type 2 diabetes	Self-reported	4670	71722	76392	93	53	28		
	Clinical records	3477	72915	76392	93	49	21	49	19
Hypertension	Self-reported	31183	83081	114264	13	10	7		
	Clinical records	26201	88063	114264	13	10	6	10	6
Stroke	Self-reported	1528	74864	76392	9	1	0		
	Clinical records	1107	75285	76392	9	0	0	0	0
Coronary heart disease	Self-reported	7516	68876	76392	46	20	5		
	Clinical records	8435	67957	76392	46	20	5	15	4

Supplementary Figure 1



Supplementary Figure 2