The Concept

The relative recurrence risk is a relative measure of recurrences of disease in relatives conditional of earlier disease has been diagnosed in the family. As such it can be used as a measure of family clustering of disease. The relative recurrence risk may be define as a ratio of two hazards λ_i/λ_j where

λ_i = probability of disease recurrence among relatives with a diseased proband

λ_j = probability of disease among relatives with no diseased proband

Thus, the relative recurrence is simply a special case of a relative risk but where the exposure is the existence of disease in a relative. This allows us to estimate the relative recurrence risk by usual tools and models for relative risk estimation including logistic regression or Cox regression. Since we have access to detailed date of onset data we are utilizing this by fitting Cox regression.

In this study we are only considering sibling and cousin relations. In this appendix I only refer to siblings but the interpretation for cousins and possibly other family relations should be straightforward.

Calculations and Statistical Model

For the estimation we are fitting Cox regression models using the years from cohort entry as time scale and using the hazard ratios from the Cox regression to estimate the relative recurrence risk risk.

The hazard being the instantaneous risk of ASD diagnosis, that is, the probability of ASD diagnosis at time t (years from cohort entry). With T indicating the random time point of ASD diagnosis this can be written as

$$\lambda_j(t)=Pr(T=t|T\geq t), j=1,2,\ldots$$

and the Cox regression can be written as $\lambda_j(t)=\lambda_0(t)e^{X_j\beta}$ where λ_0 is called the baseline hazard, a non-negative function of time of arbitrary and unspecified shape. The hazard ratio of two subjects with fixed covariate X_i and X_j is then $\frac{\lambda_i}{\lambda_j}=\frac{\lambda_0 e^{X_i\beta}}{\lambda_0 e^{X_j\beta}}=\frac{e^{X_i\beta}}{e^{X_j\beta}}=e^{(X_i-X_j)\beta}$, X_i =1 when ASD diagnosis and X_i =0 when no ASD diagnosis. The estimation assumes the relative risk (hazard ratio) is constant for all time points. This assumption is important since it allow us to remove the arbitrarily shaped baseline hazard from the comparison of exposed and non-exposed.

In our study, the units (subjects) entering the calculations are the sibling pairs in each family. Each pair consist of a child and his sibling proband. Furthermore, instead of considering each child as exposed or not exposed we allow each child to be unexposed up to the point when his proband is observed with an ASD diagnosis. The exposure is time-varying.

The eFigure 1 and eFigure 2 below illustrate how siblings enter the calculations for four different families. The eFigure 1 illustrate the siblings ordered (x-axis) by increasing birth year and the eFigure 2 show how these siblings are represented as sibling-pairs in the calculations. For each family, all pair wise sibling-pairs are included where each pair contribute with information how one sibling proband is exposing and affecting another sibling (in his family).

Thus, for the top left family (eFigure1) the first sibling is born 1992 but he do not contribute with any information about risk carried over from a sibling until his sibling is born in 1994. Now, for this family, there are two sibling pairs contributing to the calculations (eFigure2). The youngest sibling is included as being exposed by his older sibling from age 0 up to age 12, in the figure's indicated as pair (2, 1), and the older sibling is also included as being exposed by his younger sibling starting from age 2 up to age 14, in the figure's indicated as pair (1,2). Since none of these siblings experience an ASD diagnosis they will only contribute with information about the baseline risk.

The bottom left family (eFigure1) do not contribute with any information since the oldest sibling is censored (dead or emigrated) before the second sibling is born and the exposure status is not known. However, if sibling 1 had been observed with an ASD diagnosis before being censored the sibling pair (1,2) would have contributed to the calculations; but sibling pair (2,1) would still not have contributed.

The bottom right family have three children born 1992, 1995 and 1997 (eFigure1). This family will contribute with six sibling pairs (eFigure2) as each sibling can be considered exposed by each of his siblings. From 1995 sibling pairs (1,2) and (2,1) will contribute with information in the calculations of RR and from 1997 the sibling pairs (3,1) and (3,1) and (3,2) and (2,3) will contribute.

Including all pairwise combinations of siblings will have an averaging effect. Each sibling pair will enter the statistical model and the RR is calculated as the RR averaged over all the sibling pairs. Alternatively, one can consider the pairs in the family as repeated measures similar to a clinical trial with different patients contributing with different number of repeated visits or different number of blood pressure measurements.

The 'standard' Cox model assumes independence between subject. When experimental units are naturally or artificially clustered, failure times of siblings within a cluster are correlated. In the calculations we adjust for this by using a robust sandwich covariance matrix estimate (the Huber Sandwich estimator) to account for the within-family dependence[1] . The sandwich estimator is robust in the sense that the shape of the correlation matrix does not have to be specified[1].

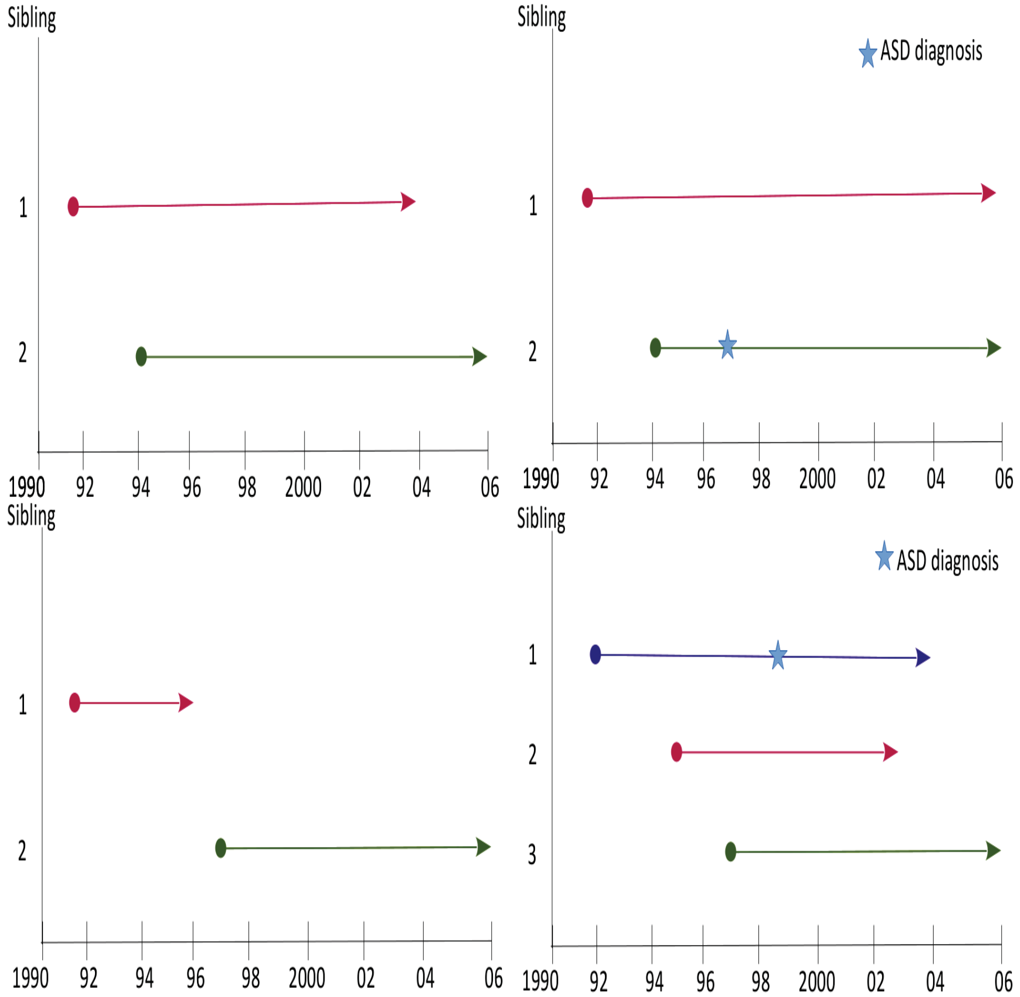
While some earlier studies have calculated the relative recurrence risk using logistic regression[2, 3] our approach is using time to event analysis which better adjust for length of follow-up where individuals followed for longer time are assumed to have higher risk than individual observed only for a short time. By adjusting for length of follow-up possible biases can be avoided and is expected to have a higher precision. Other studies have also used time to event analysis to calculate the RR[4] . Earlier studies have also been done with the additional assumptions that only younger siblings are being exposed, that is, the younger sibling become at risk for ASD when an older sibling has been diagnosed as ASD[2, 5] . Here we relax this assumption and allow any sibling to expose or be exposed by any other sibling. We do this by considering *all time-overlapping* pairwise combinations of siblings. This come with some advantages:

1. We gain power since we can utilize more events and longer follow-up
2. The measure is less restrictive since it does not need any assumption about ages between siblings.
3. With RR= $\frac{\lambda_i}{\lambda_j}=\frac{\lambda_0 e^{X_i\beta}}{\lambda_0 e^{X_j\beta}}=e^{\left(X_i-X_j\right)\beta}$ and estimating RR conditioning on older sibling always exposing a younger sibling X_j is always associated with earlier birth cohorts compared with X_i why an upward bias can potentially be introduced.e and

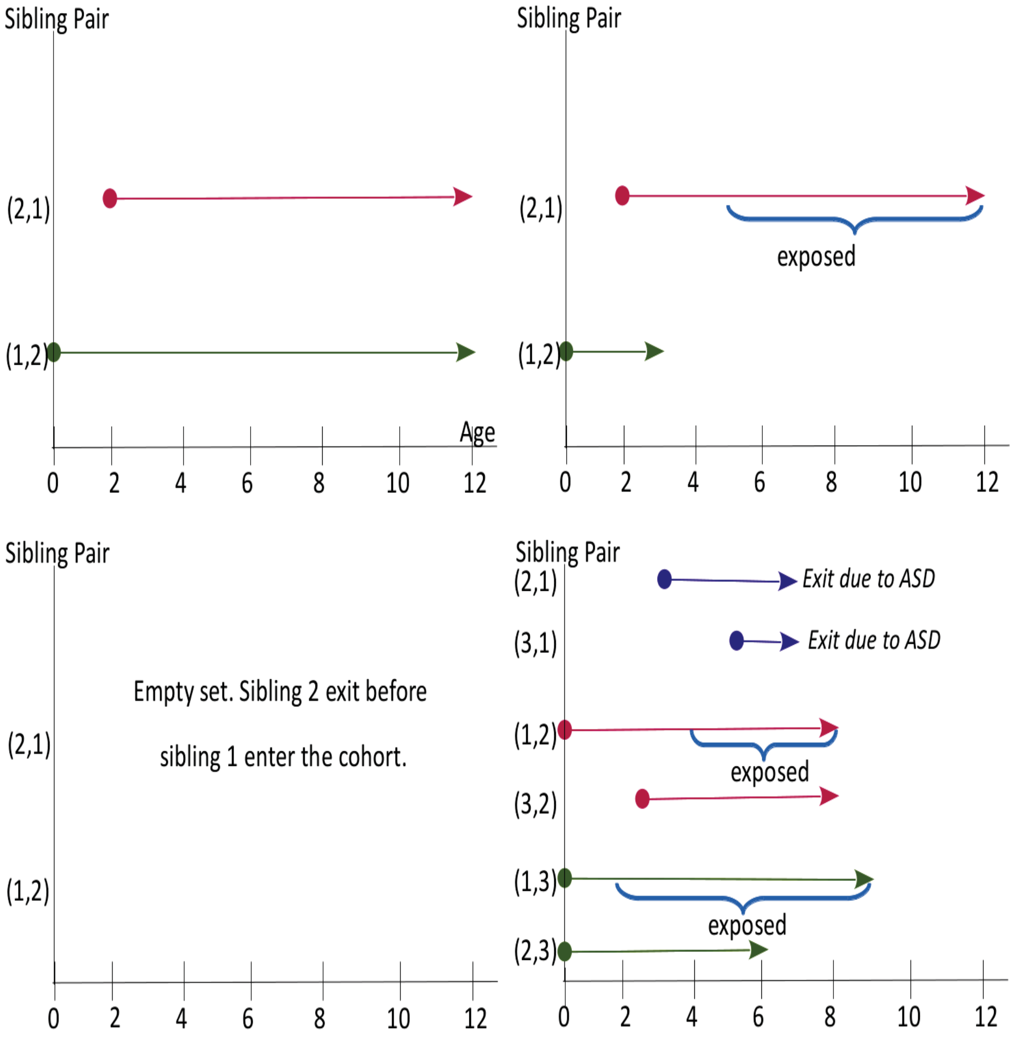
As a comparison we also calculated the RR where each sibling entering the calculations only once per family and conditioning on the older sibling exposing the later born only. We obtained the following estimates:

Family type	**Older Sibling Exposing Younger**		**Using all pair wise sibling pairs**		**Variance Ratio**
	RR (95% CI)	**Robust SE**	**RR (95% CI)**	**Robust SE**	
Full siblings	11.9 (10.6-13.3)	0.0576	10.3 (9.4-11.3)	0.0469	1.23
Maternal half siblings	3.3 (2.5-4.5)	0.1544	3.2 (2.6-4.2)	0.1245	1.24
Paternal half siblings	3.5 (2.6-4.7)	0.1573	2.9 (2.2-3.7)	0.1335	1.18

Figures - Representation of sibling-pairs



eFigure 1 Families of siblings. Example of families with siblings on calendar scale from cohort entry (solid dot) to cohort exit (death, emigration, ASD diagnosis or 31-December-2006, arrow). Sibling obtaining ASD diagnosis is marked with blue star.



eFigure 2 Sibling pairs from families in eFigure 1b as represented in the statistical analysis. Pair (1,2) represent sibling 1 exposing sibling 2 and (2,1) represent sibling 2 exposing sibling 1. Example of families with siblings on calendar scale from cohort entry (solid dot) to cohort exit (death/emigration or 31-December-2006, arrow).

Creating families family types

This section describes how the family types and subjects are defined for this study.

A diagnosis of ASD is only available from 1987 in the Swedish registers. Consequently we included all Swedish born children between 01-Jan-1982 and 31-Dec-2006 with at least one sibling. For these children essentially all parents and grandparents are known. Full siblings are siblings with the same mother and father while half-siblings share the same mother or father. Children in a half-sibling family can have siblings in full-sibling families. Twins status was found in the Swedish twin register. In the last step, using grandparents and parents id we defined cousins. In full-, half- and cousin-families with both a twin birth and a singleton birth the twins are allowed to expose a singleton sibling but not reversed.

Examples

In the example below the calculations are illustrated using R code. Software is available from http://cran.r-project.org/.

Example 1. Data representation

The families shown in eFigure 2 are represented with the follow data structure to allow calculation of the relative recurrence risk using Cox regression.

Family ID	Sibling Pair	Entry (Age)	Exit (Age)	Exposed (Y/N)	Event (Y/N)
1	2,1	2	12	0	0
1	1,2	0	12	0	0
2	2,1	2	5	0	0
2	2,1	5	12	1	0
2	1,2	0	3	0	1
4	2,1	3	7	0	1
4	3,1	5	7	0	1
4	1,2	0	4	0	0
4	1,2	4	8	1	0
4	3,2	3	8	0	0
4	1,3	0	2	0	0
4	1,3	2	9	1	0
4	2,3	0	6	0	0

Example 2. Cox regression

Below a Cox regression model is fitted for a sample with 10,000 rows and 46 events to calculate the RR using the R software.

```
library(survival); #-- Call the survival package

coxph(Surv(entry, exit, event==1) ~ asd + cluster(famid), data=A)

Call:
coxph(formula = Surv(entry, exit, event == 1) ~ asd.exp + cluster(famid),
    data = A, method = "breslow")

          coef exp(coef) se(coef) robust se    z       p
asd.exp1  3.4      29.8    0.725     0.713 4.76 1.9e-06

Likelihood ratio test=9.62  on 1 df, p=0.00192  n= 10000, number of events= 46
```

The relative recurrence risk is obtained from exp(coef) above, RR=29.8. The model standard error is 0.725 and the robust standard error, adjusting for the within-family dependencies between sibling-pairs, is 0.713.

Example 3. Number of siblings in families

Below is an example showing how family size is adjusted for in the statistical model. We multiply the number of siblings in the families in previous example with 6. Instead of having families with 2, 3 or 4 members we now have 12, 18 or 24, each member in the families replicated six times. Using the robust standard error in family clusters the RR and the standard error is unchanged. Note however how the model standard error is now considerably lower, 0.362, and deflated due to the dependent sibling-pairs.

```
#-- Replicate the earlier dataset 6 times
B = rbind(A,A,A,A)

#-- Run the code from example 1 but on the new dataset
coxph(Surv(entry, exit, event==1) ~ asd.exp + cluster(famid), data=B)


Call:
coxph(formula = Surv(entry, exit, event == 1) ~ asd.exp + cluster(famid),
    data = B, method = "breslow")

          coef exp(coef) se(coef) robust se    z       p
asd.exp1  3.4      29.8    0.362     0.713 4.76 1.9e-06

Likelihood ratio test=38.5  on 1 df, p=5.51e-10  n= 40000, number of events= 184
```

The point estimate, RR, remains the same 'exp(coef)' = 29.8. The confidence interval obtained from treating families as clusters in the family also remain unchanged.

References

1. Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. 1st ed. Wiley-Interscience; 2004

2. Lichtenstein P, Björk C, Hultman CM, Scolnick E, Sklar P, Sullivan PF. Recurrence risks for schizophrenia in a Swedish national cohort. Psychol Med. 2006;36(10):1417-1425.

3. Constantino JN, Todorov A, Hilton C, et al. Autism recurrence in half siblings: strong support for genetic mechanisms of transmission in ASD. Mol Psychiatry. 2013;18(2):137-138.

4. Grønborg TK, Schendel DE, Parner ET. Recurrence of Autism Spectrum Disorders in Full- and Half-Siblings and Trends Over Time: A Population-Based Cohort Study. *JAMA Pediatr*. 2013.

5. Ozonoff S, Young GS, Carter A, et al. Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. *Pediatrics*. 2011;128(3):e488–495.

The classical twin design

Heritability is the proportion of total variation of a disease phenotype that can be attributed to genetic sources. Historically data from twins and 'twin-models' have been standard approach for calculating heritability. These methods utilize the relation between monozygotic (MZ) and dizygotic (DZ) twins. The MZ twins in a pair are assumed to share 100% of their co-segregating genes while DZ twins are assumed to share 50%. Further, MZ and DZ twin pairs are assumed to be equally affected by shared environmental experiences, known as the equal environments assumption. From these assumptions it is possible to infer how much of the variation in a phenotype that may be attributed to genetic and environmental sources.[1]

Rational of using extended sibling design

In the current study the fraction variance explained by genetic sources (heritability) and environmental sources were estimated using an extended sibling design, an extension of the classical twin design[1]. In this design, additional to MZ and DZ twins, we included full siblings and maternal and paternal half siblings. Non-twin siblings were included in the analyses to increase power and make results more generalizable; they also allowed us to estimate both additive and dominance genetic sources of variance simultaneously as shared and non-shared environmental sources of variance.

Liability-threshold approach

For analyses of the binary outcome we employed the liability-threshold approach, much similar to probit regression. In this approach each individual is assumed to have a liability of having the disease which is distributed as the standard normal distribution, if the liability is above an estimated threshold we observe a 1 (individual has the disease), else a 0 is observed (individual does not have the disease). This may be written as

$$\Phi(1-P(Y=1))=\tau ,$$

where Φ is the distribution function of the standard normal distribution, $P(Y=1)$ is the probability of observing a 1 in the outcome, and τ is the threshold. We may further include fixed and random effects into the model;

$$\Phi(1-P(Y_k=1\vee x_k, z_k))=\tau+x_k^T\beta+z_k^T b ,$$

where k is a subject number, x_k is a vector of covariates, z_k is a design vector for random effects, β is a vector of regression coefficients, and b is a vector of random effects. In the current study x_k includes birth periods and gender, and the genetic and environmental sources of variance are defined as random effects and captured by by z_k and b .

Similarity in liabilities for disease

To test whether the liabilities for disease was similar between sibling order in a pair, and between sibling types, we performed a series of tests for both outcomes separately. We first fitted a saturated model, where we adjusted the liability of having the disease for birth period effects and for gender effects, these parameters were assumed to be the same regardless of sibling order and sibling type throughout the analyses. The liability of having the disease, adjusted for birth period and gender, was estimated separately for each sibling type as well as for sibling 1 and sibling 2 in sibling pairs (Model 1). We then fitted a model where the liability was assumed to be the same within each sibling type, regardless of order in a pair (Model 2). Finally a model where all individuals were assumed to have the same liability was fitted (Model 3). In **eTable B1** the results for ASD are shown, and in **eTable B2** the results for AD are shown. For both outcomes the order in pairs did not affect the liability of having the outcome, the liability was, however, different for the different sibling types. Therefore we allowed the liability to be different in different types of siblings when estimating the fraction of the variance explained by genetic and environmental sources.

eTable B1: Likelihood ratio tests for similarity of liability for ASD.

Model	Estimated parameters	Difference in degree of freedom	Minus 2 log likelihood	Difference in minus 2 log likelihood	p-value #
1: Saturated model	20	NA	143,903.2	NA	NA
2: Equal liability for sibling 1 and 2 in pairs	15	5	143,909.7	6.50	0.261
3: Equal liabilities for all different sibling types	8	12	145,699.9	1,796.7	<0.001

NA, not applicable. #: p-value test if model fit better than the saturated model H0: models equal Ha: model fit better

eTable B2: Likelihood ratio tests for similarity of liability for AD.

Model	Estimated parameters	Difference in degree of freedom	Minus 2 log likelihood	Difference in minus 2 log likelihood	p-value #
1: Saturated model	20	NA	64,583.8	NA	NA
2: Equal liability for sibling 1 and 2 in pairs	15	5	64,584.5	0.71	0.983
3: Equal liabilities for all different sibling types	8	12	65,123.7	539.9	<0.001

NA, not applicable. #: p-value test if model fit better than the saturated model H0: models equal Ha: model fit better

Estimating sources of variance

The extended sibling design was used to decompose the variance in liability into additive genetic factors (A) reflecting additive effects of different alleles, non-additive genetic factors (D) reflecting interaction effects between alleles at the same gene locus, shared environmental factors (C) reflecting non-genetic influences that contribute to similarity within pairs of siblings and non-shared environmental factors (E) reflecting experiences that make sibling pairs dissimilar[2]. The notation of A, D, C and E for the different components follow the standard notation in twin modelling.

MZ twins were assumed to share 100% of their A in a pair, DZ twins and full siblings were assumed to share 50% of their A, and maternal and paternal half siblings were assumed to share 25% of A in a pair. Furthermore we assumed that all sibling types included shared the C-parameter in pairs, except paternal half siblings since children tend to be living with their mother while growing up. We also assumed shared D, 100% between MZ twins, 25% between DZ twins and full siblings, and not shared between half siblings. We allowed for different liabilities of the outcome in the different sibling types, and adjusted the liability for gender and birth period.

The A, D, C, and E was treated as random effect and assumed to be normally distributed. The regression equations for sibling pair i may thus be written as

$$\begin{bmatrix}\Phi\left(1-P\left(Y_{i1}=1\right)\right)\\ \Phi\left(1-P\left(Y_{i2}=1\right)\right)\end{bmatrix}=\begin{bmatrix}\tau_i+x_{i1}^T\beta+A_{i1}+D_{i1}+C_{i1}+E_{i1}\\ \tau_i+x_{i2}^T\beta+A_{i2}+D_{i2}+C_{i2}+E_{i2}\end{bmatrix},$$

where sub index $i1$ refers to sibling 1 and $i2$ to sibling 2. Let g_{Ai} be the assumed amount of shared A, g_{Di} be the assumed amount of shared D, and m_i be the assumed amount of shared C, all for sibling pair i. The random effects can thus be stated as

$$\begin{bmatrix}A_{i1}\\ A_{i2}\end{bmatrix}\sim N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\sigma_A^2 & g_{Ai}\sigma_A^2\\ g_{Ai}\sigma_A^2 & \sigma_A^2\end{bmatrix}\right),$$

$$\begin{bmatrix}D_{i1}\\ D_{i2}\end{bmatrix}\sim N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\sigma_D^2 & g_{Di}\sigma_D^2\\ g_{Di}\sigma_D^2 & \sigma_D^2\end{bmatrix}\right),$$

$$\begin{bmatrix}C_{i1}\\ C_{i2}\end{bmatrix}\sim N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\sigma_C^2 & m_i\sigma_C^2\\ m_i\sigma_C^2 & \sigma_C^2\end{bmatrix}\right),$$

$$\begin{bmatrix}E_{i1}\\ E_{i2}\end{bmatrix}\sim N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\sigma_E^2 & 0\\ 0 & \sigma_E^2\end{bmatrix}\right).$$

The variance sources contributes to the total variance within any individual with σ_A^2 , σ_D^2 , σ_C^2 , and σ_E^2 , respectively. Since we are using the distribution function of the standard normal distribution the variance is 1, and therefore the sum of the contributions to the variance must be 1. Thus, the fraction of the total variance explained by each variance source are the estimated values for σ_A^2 , σ_D^2 , σ_C^2 , and σ_E^2 .

Heritability

The fraction of variation in a phenotype attributable to genotypic variance is referred to as the broad sense heritability. In the ADCE model $\sigma_A^2+\sigma_D^2$ is the total modelled variation due to genes, we refer to this as broad sense heritability. In contrast, the fraction of variation explained by additive genetic source of variance, σ_A^2 , is known as the narrow sense heritability[2]. In **table 2** of the main manuscript both narrow and broad sense heritability have been calculated, with 95% profile likelihood confidence intervals.

Model fitting

We fitted a model including all potential sources of variance in a model, the ADCE model. We then fitted models excluding parameters; we compared the ACE, ADE and DCE models with the full ADCE model using likelihood ratio tests. Next, sub-models where the genetic parameters, shared environmental parameter, and both these parameters are dropped (AE, DE,CE, and E models), were tested to explain the observed data and pattern of variance using as few parameters as possible. Since E, the non-shared environment source contains random error, it was not excluded in any model. These tests, and estimates of A, D, C and E fractions of variance in liability, and their confidence intervals, are shown in the main manuscript, **table 2**.

Software

For model fitting we used likelihood based techniques, as implemented in the library OpenMx[3] in the software R.[4] OpenMx allows for use of the liability-threshold approach, and for inclusion of multiple groups where co-variation between subjects vary; in our case the five different types of siblings. Given the assumptions, encoded as different co-variation between siblings pairs in the different groups, the software finds the solution for all the modeled parameters which maximizes the likelihood of observing the data.

Appendix References

1. Neale MC, Cardon LR. *Methodology for genetic studies of twins and families.* Dordrecht ; Boston: Kluwer Academic Publishers; 1992.

2. Plomin R. *Behavioral genetics.* 6 ed. New York: Worth Publishers; 2013.

3. Boker S, Neale M, Maes H, et al. OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika.* Apr 2011;76(2):306-317.

4. R Development Core Team. *R: A language and environment for statistical computing.* 2012; [computer program].