An Analysis of Sibling Correlations in Health using Latent Variable Models[*]

Timothy J. Halliday

University of Hawai'i at Mānoa and IZA[†]

Bhashkar Mazumder

Federal Reserve Bank of Chicago

November 20, 2014

Abstract

We investigate sibling correlations in health status using the Panel Study of Income Dynamics and Bayesian methods that allow us to estimate the covariance structure of a system of latent variable equations. Across a battery

[*] We would like to thank Bo Honoré, Andriy Norets and participants at the University of Hawaii Applied Micro Workshop for useful comments.

[†] Corresponding author. Address: 2424 Maile Way; 533 Saunders Hall; Honolulu, HI 96822 Tele: (808) 956- 8615 E-mail: halliday@hawaii.edu

of outcomes, we estimate that between 50% and 60% of health status can be attributed to familial or neighborhood characteristics. Taking the principal component across all outcomes, we obtain a sibling correlation of about 53%. These estimates, which are larger than previous estimates of sibling correlations in health that rely on linear models, are more in-line with sibling correlations in income and suggest that health status, like other measures of socioeconomic success, is strongly influenced by family background. Therefore, efforts to improve the circumstances of families and communities may potentially lead to improved childhood health today and also reduce future health disparities.

Key words: Sibling correlations, Intergenerational mobility, Health, Latent variable

JEL Classification: I0, I12, J0, D3, J62

1 Introduction

How important are family background and neighborhood influences in explaining health disparities? This question is increasingly salient with the rise of inequality and the growing gap in resources between families in many industrialized countries. If family and community influences during childhood play a large role then we may anticipate that health disparities are likely to grow in coming decades as rising inequality between families is manifested in adult health outcomes. Therefore policies that address the growing disparities between families may also be a form of "health policy" in that it may improve the health of the future population with implications for social safety nets. A growing literature has also linked childhood health to future economic success, *e.g.* Almond and Currie (2011), suggesting that policies that reduce health disparities may also reduce inequality in the future.

More generally, social scientists have become increasingly interested in intergenerational mobility with respect to socioeconomic status. Clearly, health is an important component of socioeconomic status but intergenerational influences on health have been much less studied than other key measures of status such as income, education and occupation.

As an empirical matter, it is very challenging to measure the importance of family background on health. One important issue is how exactly to measure family background. A small but notable literature has used sibling correlations as a catch-all measure of family background intended to capture all influences shared in common by siblings. This sidesteps the difficulty of having to measure each of the multitude of possible measures of family background –many of which may be unavailable in most datasets. Indeed, Bjorklund and Jantti (2012) emphasize that sibling correlations are in general, much more useful than the traditional measures of intergenerational associations for studying mobility across generations.

A second critical issue is how exactly to measure health. Standard datasets with health outcomes typically contain dichotomous measures (*e.g.* asthma, disability, etc.) that might occur with low frequency. Alternatively, surveys sometimes collect relatively blunt measures such as self reported health status on a categorical scale. How can one best use such measures to get at a more ideal concept of underlying or latent health status? This paper develops the econometric tools that can be used to estimate sibling correlations in health that overcome some of the limitations encountered in previous work that, for example, uses linear models in a situation where they clearly are not appropriate.

Specifically, we consider the inter-generational transmission of health status by estimating sibling correlations in a battery of health measurements for children in the Child Development Supplement of the Panel Study of Income Dynamics (PSID-

CDS). Each of these measurements is modeled as being determined by a latent variable. The arbitrary covariance structures for the individual- and family-specific random effects are estimated using Bayesian methods. To account for the possibility that the measurements are proxies for a more fundamental latent health variable, we also estimate sibling correlations in the principal components of the covariance matrices for the two types of random effects.

Our estimates indicate that sibling correlations for a variety of health measures range between 0.5 and 0.6 with few exceptions. This suggests that over half of a child's health status can be attributed to familial or community influences. These estimates are substantially larger than those from Mazumder (2011) who uses linear models to estimate sibling correlations in health also using the PSID-CDS; his estimates for health outcomes tend to be on the order of 0.1-0.2. Notably, our estimates of sibling correlations in health are more in-line with estimates of sibling correlations in income from Mazumder (2008) who obtains an estimate of about 0.5. Finally, the sibling correlation is larger for boys than it is for girls suggesting that community or family influences matter more for boys.

Looking across all of the health outcomes using principal components analysis, we obtain a sibling correlation of 0.531 which tends to be lower than when we consider only a single health outcome. Here we can draw an analogy to Spearman's G-factor for intelligence where a single incorrect response on an exam does not necessarily indicate poor intellectual capacity, overall. Similarly, we may think of our principal component as a measure of general health status so that a high sibling correlation in one particular outcome (*e.g.* anemia) does not necessarily indicate a high sibling correlation in overall health status. On the whole, our results indicate that the role of family and community influences on health status is large and on par with their role in determining economic status.

The rest of the paper proceeds as follows. Section 2 describes the structure of our model using a SUR framework. Section 3 describes the estimation. Section 4 describes how we construct our sibling correlation measures. Section 5 describes the PSID-CDS data. , Section 6 presents the key results. In section 7, we conclude.

2 A SUR Model of Sibling Correlations

We consider a set of M binary measures of health which we will index with $m \in \{1, ..., M\}$. We observe these measures for sibling $i \in \{1, ..., N_f\}$ in family $f \in \{1, ..., F\}$ at time $t \in \{1, ..., T_m\}$.[1] We denote the mth measure for individual i in

[1] We subscripted T with m to denote that different measurements are observed for differing lengths of time.

family f at time t with h^m_{ift}. Each outcome is determined by a latent variable

$$h^{*m}_{ift} = x_{ift}\boldsymbol{\beta}^m + \alpha^m_f + \gamma^m_{if} + u^m_{ift} \tag{1}$$

where

$$h^m_{ift} = 1\left(h^{*m}_{ift} > 0\right). \tag{2}$$

The first term on the right-hand side, x_{ift}, is $1 \times K$ vector of observable heterogeneity. Because most observable variables such as parental characteristics do not change over time, we only include age and a constant in x_{ift}. The term, α^m_f, is a family-specific effect. Next, γ^m_{if}, is an individual-specific effect. Neither of these varies with time. The final component is a time-variant idiosyncratic residual. Each of these components is specific to a particular measurement and, hence, superscripted m.

We define $\boldsymbol{\alpha}_f \equiv \left(\alpha^1_f, ..., \alpha^M_f\right)'$, $\boldsymbol{\gamma}_{if} \equiv \left(\gamma^1_{if}, ..., \gamma^M_{if}\right)'$ and $\mathbf{u}_{if} \equiv (u^1_{if1}, ..., u^1_{ifT_1}, ..., u^M_{if1}, ..., u^{M\prime}_{ifT_M})'$. Note that $\boldsymbol{\alpha}_f$ and $\boldsymbol{\gamma}_{if}$ are $M \times 1$ and $\mathbf{u}_{if}$ is $T \times 1$ where $T \equiv \sum_{m=1}^{M} T_m$. In practice, T can change across individuals and families but we do not notate this to economize in the exposition. Next, we assume that

$$\begin{pmatrix} \boldsymbol{\alpha}_f \\ \boldsymbol{\gamma}_{if} \\ \mathbf{u}_{if} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0}_M \\ \mathbf{0}_M \\ \mathbf{0}_T \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0}_{M,M} & \mathbf{0}_{M,T} \\ \mathbf{0}_{M,M} & \boldsymbol{\Omega} & \mathbf{0}_{M,T} \\ \mathbf{0}_{T,M} & \mathbf{0}_{T,M} & \mathbf{I}_T \end{pmatrix}\right). \tag{3}$$

We normalize the variances of the idiosyncratic components to unity and leave $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ unrestricted.

It is useful to write this system as a SUR model in the latent variable. Defining $\mathbf{H}^*_{if} \equiv \left(h^{*1}_{if1}, ..., h^{*1}_{ifT_1}, ..., h^{*M}_{if1}, ..., h^{*M}_{ifT_M}\right)'$ and $\mathbf{x}^m_{if} \equiv \left(x'_{if1}, ..., x'_{ifT_m}\right)'$, we can write

$$\mathbf{H}^*_{if} = \begin{pmatrix} \mathbf{x}^1_{if} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \mathbf{x}^M_{if} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_M \end{pmatrix} + \begin{pmatrix} 1_{T_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1_{T_M} \end{pmatrix} \boldsymbol{\gamma}_{if} + \begin{pmatrix} 1_{T_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1_{T_M} \end{pmatrix} \boldsymbol{\alpha}_f + \mathbf{u}_{if}$$

where 1_J is a J-vector of ones. We can write this more compactly as

$$\mathbf{H}^*_{if} = \mathbf{X}_{if}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\gamma}_{if} + \mathbf{P}\boldsymbol{\alpha}_f + \mathbf{u}_{if}$$

where $\mathbf{H}^*_{if}$ and $\mathbf{u}_{if}$ are $T \times 1$, $\mathbf{X}_{if}$ is $T \times KM$ and $\mathbf{P}$ is $T \times M$. If we define $\mathbf{H}^*_f \equiv \left(\mathbf{H}^{*\prime}_{1f}, ..., \mathbf{H}^{*\prime}_{N_f f}\right)'$ and $\mathbf{u}_f \equiv \left(\mathbf{u}'_{1f}, ..., \mathbf{u}'_{N_f f}\right)'$ which are both $TN_f \times 1$, then we can write

$$\mathbf{H}_f^* = \mathbf{X}_f\boldsymbol{\beta} + \left[\mathbf{I}_{N_f} \otimes \mathbf{P}\right]\boldsymbol{\gamma}_f + \left[1_{N_f} \otimes \mathbf{P}\right]\boldsymbol{\alpha}_f + \mathbf{u}_f$$

where $\mathbf{X}_f \equiv \left(\mathbf{X}'_{1f}, ..., \mathbf{X}'_{N_f}\right)'$ which is $TN_f \times KM$ and $\boldsymbol{\gamma}_f \equiv \left(\boldsymbol{\gamma}'_{1f}, ..., \boldsymbol{\gamma}'_{N_f f}\right)'$ which is $MN_f \times 1$. Note that the second term on the right-hand side contains the identity matrix whereas the third term contains a vector of ones.

Finally, we define $N \equiv \sum_{f=1}^{F} N_f$ and stack one more time over families to obtain the full SUR system. Defining

$$\mathbf{H}^* \equiv \underbrace{(\mathbf{H}_1^{*\prime}, ..., \mathbf{H}_F^{*\prime})'}_{TN\times 1},$$

$$\mathbf{X} \equiv \underbrace{(\mathbf{X}'_1, ..., \mathbf{X}'_F)'}_{TN\times KM},$$

$$\mathbf{G} \equiv \underbrace{\mathbf{I}_N \otimes \mathbf{P}}_{TN\times NM},$$

$$\boldsymbol{\gamma} \equiv \underbrace{(\boldsymbol{\gamma}'_1, ..., \boldsymbol{\gamma}'_F)'}_{NM\times 1},$$

$$\mathbf{A} \equiv \underbrace{\begin{pmatrix} 1_{N_1} \otimes \mathbf{P} & & \mathbf{0}_{TN_1\times M} \\ & \ddots & \\ \mathbf{0}_{TN_F\times M} & & 1_{N_F} \otimes \mathbf{P} \end{pmatrix}}_{TN\times FM},$$

$$\boldsymbol{\alpha} \equiv \underbrace{(\boldsymbol{\alpha}_1', ..., \boldsymbol{\alpha}_F')'}_{FM\times 1}$$

and

$$\mathbf{u} \equiv \underbrace{(\mathbf{u}_1', ..., \mathbf{u}_F')'}_{TN\times 1},$$

we can write

$$\mathbf{H}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \mathbf{A}\boldsymbol{\alpha} + \mathbf{u}. \tag{4}$$

The task ahead will be to employ methods to estimate and conduct inference on $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ and their roots.[2]

3 Bayesian Inference

The posterior distribution of the model's parameters will be of the form

$$p\left(\boldsymbol{\Sigma}, \boldsymbol{\Omega}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathbf{H}^* | \mathbf{H}, \mathbf{W}\right) \tag{5}$$

where $\mathbf{W} \equiv [\mathbf{X}, \mathbf{A}, \mathbf{G}]$. This posterior distribution has two important features. The first is that, because the latent variable $\mathbf{H}^*$ is unobserved by the econometrician, it must be simulated. This can easily be done within the Gibbs sampler by employing

[2] In the model as written, we have a constant, NM individual fixed effects and FM family fixed effects which are not separately identified. So, in what proceeds, we will estimate $(F-1)\,M$ family effects and $(N-F)\,M$ individual effects.

the data augmentation procedure first described by Albert and Chib (1993). The second is that once we have simulated the latent variable, we can condition on it as if it was data.

Conditional on $\mathbf{H}^*$, the model will then have the following hierarchical structure:

$$\begin{aligned} p\left(\boldsymbol{\Sigma}, \boldsymbol{\Omega}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha} | \mathbf{H}^*, \mathbf{W}\right) \propto\ & p\left(\mathbf{H}^* | \boldsymbol{\Sigma}, \boldsymbol{\Omega}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathbf{W}\right) \times \\ & \prod_{f=1}^{F} \prod_{i=1}^{N_f - 1} p\left(\boldsymbol{\gamma}_{if} | \boldsymbol{\Omega}\right) \prod_{f=1}^{F-1} p\left(\boldsymbol{\alpha}_f | \boldsymbol{\Sigma}\right) \times \\ & p\left(\boldsymbol{\beta}\right) p\left(\boldsymbol{\Sigma}\right) p\left(\boldsymbol{\Omega}\right). \end{aligned}$$

The first term on the right-hand side is the likelihood of the latent variable which is, in fact, the likelihood for the Classical Fixed Effects model in the latent variable. The second term is the prior on the family and individual specific fixed-effects in equation (1) and is given by the distribution in equation (3). The final term includes the priors on $\boldsymbol{\Sigma}, \boldsymbol{\Omega}$ and $\boldsymbol{\beta}$. We use the following conjugate priors:

$$\begin{aligned} & \boldsymbol{\beta} \sim N\left(\mathbf{0}, \underline{\mathbf{H}}_{\beta}^{-1}\right) \\ & \boldsymbol{\Sigma} \sim IW(\underline{\mathbf{S}}, v) \\ & \boldsymbol{\Omega} \sim IW\left(\underline{\mathbf{V}}, w\right). \end{aligned}$$

Note that the second term is conditional on $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ and that the terms, $p\left(\boldsymbol{\Sigma}\right)$ and $p\left(\boldsymbol{\Omega}\right)$, are, in fact, priors on priors or hyperpriors.[3]

Since conditioning on the latent variable reduces the model to a standard hierarchical linear model or the variance-components model discussed in Browne and Draper (2006), we can easily estimate it using the Gibbs sampler. We will proceed in a series of steps. Before we delineate these, we will first discuss some key conditional distributions that will be needed to implement the Gibbs sampler. First, we will sample $\boldsymbol{\Psi} \equiv [\boldsymbol{\beta}', \boldsymbol{\alpha}', \boldsymbol{\gamma}']'$ so that the regression coefficients and the fixed effects are sampled as a single block.[4] The conditional distribution for $\boldsymbol{\Psi}$ is then given by

$$\boldsymbol{\Psi}|\boldsymbol{\Sigma}, \boldsymbol{\Omega}, \mathbf{X}, \mathbf{H}^{*} \sim N\left(\boldsymbol{\Psi}_{*}, \mathbf{H}_{\boldsymbol{\Psi}}^{-1}\right) \tag{6}$$

[3] Because we have a large number of families in our data, the choice of the prior is not terribly important as it is well known that the posterior and likelihood functions become closer together as the sample size increases (see Theorems 3.4.2 and 3.4.3 from Geweke (2005)).

[4] In practice, we also experimented with sampling $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ separately. It turned out that this was slightly faster because it required the inversion of smaller matrices. So, this is what we used although to conserve on notation, we only describe sampling them as a single block in the manuscript. However, the basic ideas are the same in either case.

where

$$\begin{aligned}
\mathbf{H}_{\boldsymbol{\Psi}} &\equiv \mathbf{W}'\mathbf{W}+\underline{\mathbf{H}}_{\boldsymbol{\Psi}} \\
\underline{\mathbf{H}}_{\boldsymbol{\Psi}} &\equiv diag(\underline{\mathbf{H}}_{\beta}, \boldsymbol{\Sigma}^{-1}, \boldsymbol{\Omega}^{-1}) \\
\widehat{\boldsymbol{\Psi}} &\equiv \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\mathbf{H}^{*} \\
\boldsymbol{\Psi}_{*} &\equiv \mathbf{H}_{\boldsymbol{\Psi}}^{-1}\mathbf{W}'\mathbf{W}\widehat{\boldsymbol{\Psi}}.
\end{aligned}$$

Now that we have sampled $(\boldsymbol{\beta}', \boldsymbol{\alpha}', \boldsymbol{\gamma}')'$, we can sample from the conditional posterior of $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma}|\boldsymbol{\Omega}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma} \sim IW\left(\boldsymbol{\Sigma}_{*}, v_{*}\right) \tag{7}$$

where

$$\begin{aligned}
\boldsymbol{\Sigma}_{*} &\equiv \sum_{f=1}^{F-1}\boldsymbol{\alpha}_{f}\boldsymbol{\alpha}_{f}' + \underline{\mathbf{S}} \\
v_{*} &= F - M - 2 + v.
\end{aligned}$$

Similarly, the conditional posterior of $\boldsymbol{\Omega}$ is

$$\boldsymbol{\Omega}|\boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma} \sim IW\left(\boldsymbol{\Omega}_{*}, w_{*}\right) \tag{8}$$

where

$$\boldsymbol{\Omega}_* \equiv \sum_{f=1}^{F} \sum_{i=1}^{N_f - 1} \boldsymbol{\gamma}_{if} \boldsymbol{\gamma}'_{if} + \underline{\mathbf{V}}$$

$$w_* \equiv N - F - M - 1 + w.$$

To sample from the posterior in (5), we will sample from these conditional distributions inside of the Gibbs sampler.[5] This will work in the following steps.[6]

1. Initialize $\left(\boldsymbol{\Sigma}^0, \boldsymbol{\Omega}^0, \boldsymbol{\beta}^0, \boldsymbol{\gamma}^0, \boldsymbol{\alpha}^0\right)$.

2. Sample from $p\left(\mathbf{H}^* | \boldsymbol{\Sigma}^{n-1}, \boldsymbol{\Omega}^{n-1}, \boldsymbol{\beta}^{n-1}, \boldsymbol{\gamma}^{n-1}, \boldsymbol{\alpha}^{n-1}, \mathbf{H}, \mathbf{W}\right)$. Specifically, draw TN values of $\mathbf{H}^*$ from the conditional distribution which will be a truncated Normal distribution. Once these are drawn, they should be treated as data. This is the data augmentation step.

[5] To see why this is the number of degrees of freedom, note that the part of the conditional posterior for $\boldsymbol{\Sigma}$ ignoring the prior is

$$|\boldsymbol{\Sigma}|^{-\frac{1}{2}(F-1)} \exp\left(tr \left(\boldsymbol{\Sigma}^{-1} \sum_{f=1}^{F-1} \boldsymbol{\alpha}_f \boldsymbol{\alpha}'_f \right)\right).$$

If look at the definition of the Inverted Wishart from p. 305 of Bauwens, Lubrano, and Richard (1999), we see that

$$F - 1 = dof + M + 1$$

so that the degrees of freedom coming from this portion of the posterior must be $F - M - 2$. The calculation for $\boldsymbol{\Omega}$ is similar.

[6] We conducted Monte Carlo experiments and found no issues with our Bayesian estimation procedure. The Gibbs' sampler that we employed converged to a variety of hypothetical data generating processes. In addition, we also experimented with a number of starting values for the Gibbs' sampler and this did not affect convergence.

3. Sample from $p\left(\boldsymbol{\Psi}|\mathbf{H}^{*n},\boldsymbol{\Sigma}^{n-1},\boldsymbol{\Omega}^{n-1},\mathbf{H},\mathbf{W}\right)$ using the distribution in (6).

4. Sample from

$$p\left(\boldsymbol{\Omega}|\boldsymbol{\gamma}^{n},\boldsymbol{\alpha}^{n},\boldsymbol{\beta}^{n},\mathbf{H}^{*n},\boldsymbol{\Sigma}^{n-1},\mathbf{H},\mathbf{W}\right)$$
$$p\left(\boldsymbol{\Sigma}|\boldsymbol{\Omega}^{n},\boldsymbol{\gamma}^{n},\boldsymbol{\alpha}^{n},\boldsymbol{\beta}^{n},\mathbf{H}^{*n},\mathbf{H},\mathbf{W}\right)$$

   using the distributions in (7) and (8).

5. Go back to step 2 and repeat.

4 Measuring Sibling Correlations

We propose two ways of measuring sibling correlations. The first is the most straight forward. For each of the M health measurements, we sample

$$\rho^{m}\equiv\frac{\sigma_{\alpha}^{m}}{\sigma_{\gamma}^{m}+\sigma_{\alpha}^{m}} \tag{9}$$

and conduct inference on the correlation for the mth health measurement. However, an alternative is to view the M different health measures as proxies for a latent health variable. So, ultimately, we may not care about the intra-household correlation in any given measure *e.g.* ρ^m , but rather the sibling correlation in some broader measure of latent health.

For the second sibling correlation, we will require some way of reducing the information in the matrices $\mathbf{\Sigma}$ and $\mathbf{\Omega}$, so that we can, essentially, operationalize the notion of "$\frac{\mathbf{\Sigma}}{\mathbf{\Sigma}+\mathbf{\Omega}}$" into a single correlation. Probably, the most common way of reducing the information in the vectors $\boldsymbol{\alpha}_f$ and $\boldsymbol{\gamma}_{if}$ is to conduct a Principal Components Analysis (PCA) of $\mathbf{\Sigma}$ and $\mathbf{\Omega}$. We can then compute sibling correlations based on these components.

To fix ideas, we denote these eigenvalues by $\delta_1, ..., \delta_M$ and $\lambda_1, ..., \lambda_M$ of $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ in descending order. We define these new intra-household correlations that pool information across health outcomes as

$$\pi^\mu = \frac{\sum_{m=1}^{\mu} \delta_m}{\sum_{m=1}^{\mu} \delta_m + \sum_{m=1}^{\mu} \lambda_m}. \tag{10}$$

So, if $\mu = 1$ then we consider the sibling correlation in only the first principal component. If $\mu = M$ then we consider all the components and so $\pi^M = \frac{tr\left(\mathbf{\Sigma}\right)}{tr\left(\mathbf{\Sigma}\right) + tr\left(\mathbf{\Omega}\right)}$. For each draw of the matrices, $\mathbf{\Omega}$ and $\mathbf{\Sigma}$, from the posterior, we will compute π^μ to obtain the posterior distribution of our measure of the sibling correlation.

5 Data

As discussed above, we employ the PSID-CDS on children 18 years of age or younger. The data come from the years 1997, 2002/2003 and 2007/2008. We used the PSID-CDS to measure a battery of health outcomes which are listed in Table 1 together with their descriptive statistics. These measures are binary indicators for various conditions, disabilities or other outcomes pertinent to a child's health. Most of them are self-explanatory except for Self-Reported Health Status (SRHS) which is a categorical variable in which the respondent classified her own health as excellent (SRHS = 1), very good (SRHS = 2), good (SRHS = 3), fair (SRHS = 4) and poor (SRHS = 5). In our analysis, we will break the SRHS measure into three dummy variables indicating SRHS greater than or equal to 2,3 or 4. As discussed above, the first stage of the hierarchical model is essentially a Classical Fixed Effects model and so, there is no need to include time invariant characteristics in it. As such, the only explanatory variable in the model (other than a constant) is age and its descriptive statistics are reported in Table 2. Finally, we also estimate the model for certain subsets of the data. For these, we stratify the data by the average of parental income over the child's duration in the sample or by gender.

In Table 3, we report the number of observations that we have for each of our 10 measurements for the first, second and third years present in the sample. For the first year that the respondent was present, which we call the baseline, we have 3235 observations. Note that the first year present need not be 1997 since many children in our data were born after this year. In total, we have data on 3235 individuals in 2173 households.

6 Results

6.1 Checking Convergence

We ran the Gibbs sampler for 20,000 iterations. To gauge the convergence of the sampler, we employed the CUMSUM statistic from Yu and Myckland (1998) which is given by

$$CUMSUM_t = \left(\frac{1}{t} \sum_{n=1}^{t} \theta^n - \mu_\theta \right) / \sigma_\theta$$

where μ_θ and σ_θ are the mean and the standard deviation for all 20,000 iterations. If the sampler converges to a stationary distribution then $CUMSUM_t$ will converge smoothly to zero. We report the $CUMSUM_t$ statistics in Figure 1 for the elements of $\boldsymbol{\beta}$ and the diagonal elements of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$. The figures show a smooth convergence towards zero as should be the case if the sampler converges. To account for the "burn-in" phase of the sampler in which it is still converging, for the coming results, we do not use the first 1000 iterations which this figure indicates may be a bit off from the limiting distribution. In Figure 2, we report the time series for all 20,000 iterations for the diagonal elements of $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ and the three highest components of the corresponding covariance matrices. The figure reveals that, from an early point in the sampler, the distribution is stationary.

Finally, in Table 4, we estimate an AR(1) model of the form

$$z_t = \phi_0 + \phi_1 z_{t-1} + e_t$$

where z_t represents a sampled parameter at iteration t using OLS and computing Newey-West standard errors. We estimate ϕ_1 for the diagonal elements of $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ as well as for the sibling correlations, ρ. The results indicate that all of the estimates of ϕ_1 are all significantly below one but do indicate a fair amount of persistence in the sampled parameters.

6.2 Core Results

Our core results can be found in Table 5 where we report the mean, median and standard deviation of the sibling correlations defined in equation (9) for each of the twelve outcomes that we consider. We also report the sibling correlations that are based on the principal components defined in equation (10) at the bottom of the table. To provide the reader with a visual idea of the distribution of these correlations, we plot their distributions in Figure 3 using box plots.

The table and the figure reveal that the sibling correlations for all the outcomes tend to be between 0.45 and 0.75 indicating that at least half of a child's latent health can be attributed to familial or environmental circumstances. The medians and the means are virtually identical indicating that the distribution of the correlations is highly symmetric.

When we look at the correlations based on the components of the covariance matrices at the bottom of the table, we see that they are smaller than for any one outcome; they are now between 53% and 57%. Perhaps this is not surprising since these correlations reflect a deeper notion of health status. Just because a sibling pair has a high propensity for experiencing a particular outcome does not imply that they have a similarly high propensity for experiencing all of the outcomes that we consider which suggests that the correlation in the principal components should be smaller.

6.3 Demographic Subsets

We also estimated the model for certain demographic subsets. The results by gender are reported in Table 6 and Figure 4. On the whole from looking at the table, it is hard to tell if the correlations are higher for boys or girls. However, looking at the correlation based on the first principal component, π^1(which may be viewed as the best summary of the available information), we do see that the correlation for girls is 0.540 whereas it is 0.624 for boys. A formal test that the mean of the sibling correlations for boys and girls is different that utilizes Newey-West standard errors indicates that these difference are indeed statistically significant ($p < 0.001$).[7]

We also estimated the model by parental income quartile. We do not report these results to save space but they are available in an on-line appendix. On the whole, these results did not turn up any salient patterns.

6.4 REML Estimates

We now present a set of estimates of sibling correlations from our data using Restricted Maximum Likelihood (REML) which has been commonly used in the literature.[8] Specifically, we estimate a model of the form

$$h_{ift}^{m} = x_{ift}\boldsymbol{\beta}^{m} + \alpha_{f}^{m} + \gamma_{if}^{m} + u_{ift}^{m} \tag{11}$$

which is a linear version of the model that we have considered throughout the paper. The vector x_{ift} now includes a constant, age and sex.

[7] Note that Table 6 reports the standard deviations not standard errors.

[8] See Mazumder (2008), Björklund, Lindahl, and Lindquist (2010), Mazumder (2011) and Schnitzlein (2014).

In Table 7, we report estimates of ρ^m from this linear model. The estimates from the linear model are smaller than those from the non-linear model for six of the nine measures excluding the SRHS variables. For example, we obtain a sibling correlation of 0.277 for asthma from model (11) and an estimate of 0.486 from model (1). Similarly, for diabetes, we obtain 0.209 from the linear model and 0.628 from the latent variable model. Of these nine measures, the only REML estimates that are larger are for anemia, allergies and limitations on school attendance. On the other hand, the estimates for SRHS are larger in the linear model than in the latent variable model. On the whole, it appears as if the estimates from the latent variable model in Table 5 are less variable in that they tend to hover between 0.5 and 0.6, whereas the REML estimates in Table 7 range from 0.165 to 0.930. It is also noteworthy that the classical confidence intervals in Table 7 which are based on a Normal approximation of the finite sample distribution often contain unity which is a pathology that is not present with Bayesian confidence intervals.

In summary, our estimates from the latent variable model paint a much more accurate picture of the intergenerational transmission of health status.[9] In the Appendix, we show the results of a simple Monte Carlo exercise in which REML

[9] The estimates in Table 7 tend to be higher than those in Mazumder (2011) who uses the same data. The reason for this is that, for many of the variables, Mazumder (2011) uses a variable for having "ever had" the condition that eliminates the time dimension of these variables.

estimates which assume that the measurements in equation (2) are linear are severely biased. Based on this, we conclude that properly modeling the latent variable is crucial when estimating sibling correlations in health.

7 Conclusion

In this paper, we investigate the role of family background and community influences in explaining health disparities which is a topic that has received scant attention in the literature. Using the CDS of the PSID, we estimate sibling correlations across a battery of health outcomes that are on the order of 0.5 to 0.6 and these appear to be higher for boys than for girls. If we consider the principal component across all of the measurements, which can be viewed as akin to the G-factor for intelligence, we obtain a correlation of 0.531. These findings suggest that at least 50% of the variation in children's health can be attributed to family or community influences which is larger than previous estimates of sibling correlations in health from Mazumder (2011) and more in line with the estimated sibling correlation in adult wages in the US found by Mazumder (2008). Importantly, as argued by Bjorklund and Jantti (2012), the sibling correlation should be viewed as a lower bound of the importance of family background as there are many important family characteristics that are not shared by siblings. This suggests that policies that can reduce disparities in resources between families and communities can potentially reduce inequality in childhood health today as well as future disparities in adult health.

There is also a growing literature that shows that improved health early in life can have lasting effects on economic outcomes later in life *e.g.* Almond and Currie (2011). This suggests that efforts to reduce childhood health disparities may also reduce inequality in the future thereby attenuating the transmission of economic status across generations.

Future research may wish to better understand the precise mechanisms that underpin the sizable sibling correlations in health. For example, how important are neighborhood influences such as peers and schools compared to family characteristics such as income and parental education. A better understanding of the sources of the sizable sibling correlation in health can provide useful information to guide policy makers in their efforts to reduce health disparities in the population.

8 Appendix

We report the results of a Monte Carlo exercise. We simulated the model in equations (1) and (2) for $M = 2$, $F = 2000$, $N = 2$ and $T = 3$. As above, we only included a constant and a time trend. We employed the following parameter values:

$$\mathbf{\Omega} = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1.5 \end{bmatrix}; \mathbf{\Sigma} = \begin{bmatrix} 2 & 0.1 \\ 0.1 & 2 \end{bmatrix}; \beta = (0.2, -2.5, 0.2, -3)'.$$

These parameter values imply that $\rho^1 = 0.67$ and $\rho^2 = 0.57$. We generated 5 samples using this data generating process (DGP). We then used REML to estimate two models. The first was the model in equation (1) so that we assumed that we observed the latent variable; we call this model LV. These estimates should be close to those from the true DGP. The second model erroneously assumed that the measurements were linear as in equation (11); we call this model M.

The results are reported in Tables 8 and 9 for $m = 1$ and $m = 2$, respectively. Not surprisingly, model LV delivers the correct answer in all cases. In contrast, model M delivers very biased estimates. The estimates of the variance components are substantially biased. For example, in Table 8, the estimates of σ^2_α are about 0.15% of the true values from the DGP. For the misspecified model M, the estimates of ρ^m for $m = 1, 2$ are slightly more accurate but still severely biased. In Table 8, the estimates of ρ^1 range from 0.564 to 0.623 when the value implied by the DGP is 0.667. In Table 9, the estimates of ρ^2 range from 0.410 to 0.458 when the DGP implies a true value of 0.571.

References

Albert, J., and S. Chib (1993): "Bayesian Analysis of Binary and Polychotomous Data," *Journal of the American Statistical Association*, 88(422), 669–679.

Almond, D., and J. Currie (2011): "Human Capital Development Before Age 5," in *Handbook of Labor Economics vol. 4*, ed. by D. Card, and O. Ashenfelter. North-Holland, Amsterdam.

Bauwens, L., M. Lubrano, and J.-F. Richard (1999): *Bayesian Inference in Dynamic Econometric Models*. Oxford, Oxford, UK.

Bjorklund, A., and M. Jantti (2012): "How Important is Family Background for Labour-economic Outcomes?," *Labour Economics*, 19(4), 465–474.

Björklund, A., L. Lindahl, and M. J. Lindquist (2010): "What more than parental income, education and occupation? An exploration of what Swedish siblings get from their parents," *The BE Journal of Economic Analysis & Policy*, 10(1).

Browne, W., and D. Draper (2006): "A Comparison of Bayesian and Likelihood-Based Methods for Fitting Multilevel Models," *Bayesian Analysis*, 1(3), 473–514.

GEWEKE, J. (2005): *Contemporary Bayesian Econometrics and Statistics*. Wiley, Hoboken, New Jersey.

MAZUMDER, B. (2008): "Sibling Similarities and Economic Inequality in the U.S.," *Journal of Population Economics*, 21(3), 685–701.

——— (2011): "Family and Community Influences on Health and Socioeconomic Status: Sibling Correlations Over the Life Course," *B.E. Journal of Economic Analysis and Policy (Contributions)*, 11(3).

SCHNITZLEIN, D. D. (2014): "How important is the family? Evidence from sibling correlations in permanent earnings in the USA, Germany, and Denmark," *Journal of Population Economics*, 27(1), 69–89.

YU, B., AND P. MYCKLAND (1998): "Looking at Markov Samplers Through CUMSUM Path Plots: A Simple Diagnostic Idea," *Statistics and Computing*, 8(3), 275–286.

Table 1: Health Outcomes

	Mean (SD)
Asthma	0.139 (0.346)
Diabetes	0.004 (0.059)
Anemia	0.052 (0.223)
Development Delay	0.052 (0.223)
Hyperactivity	0.064 (0.225)
Allergies	0.161 (0.368)
Limitations on Athletics	0.042 (0.200)
Limitations on School Attendance	0.018 (0.132)
Limitations on School Work	0.032 (0.176)
Self-Reported Health Status*	1.675 (0.813)

*Denotes 5-point categorical variable. All others are binary.

Table 2: Exogenous Covariates

	Mean (SD)
Age - 1st Year Present	6.178 (3.632)
Age - 2nd Year Present	11.125 (3.696)
Age - 3rd Year Present	13.486 (2.198)

Table 3: Sample Sizes by Measurement

Health Measure	1st Year Present	2nd Year Present	3rd Year Present
Asthma	3235	2783	1344
Diabetes	3235	2784	1345
Anemia	3235	2785	1344
Development Delay	2325	2785	1344
Hyperactivity	3235	2782	1339
Allergies	3235	2789	1345
Limitations on Athletics	3235	2785	1345
Limitations on School Att.	3235	2784	1339
Limitations on School Work	3235	2784	1333
SRHS	3235	2780	1337

Table 4: Autogression Coefficients for Variance Components and Sibling Correlations

	Omega	Sigma	Rho
Asthma	0.908 (0.003)	0.933 (0.002)	0.824 (0.004)
Diabetes	0.985 (0.001)	0.989 (0.001)	0.980 (0.002)
Anemia	0.979 (0.002)	0.971 (0.002)	0.969 (0.002)
Development Delay	0.930 (0.002)	0.951 (0.002)	0.875 (0.003)
Hyperactivity	0.957 (0.002)	0.938 (0.002)	0.914 (0.003)
Allergies	0.946 (0.003)	0.929 (0.002)	0.918 (0.003)
Limitations on Athletics	0.941 (0.002)	0.949 (0.002)	0.901 (0.003)
Limitations on School Att.	0.970 (0.002)	0.982 (0.001)	0.948 (0.002)
Limitations on School Work	0.944 (0.002)	0.950 (0.002)	0.908 (0.003)
SRHS >= very good	0.936 (0.002)	0.885 (0.003)	0.907 (0.003)
SRHS >= good	0.950 (0.002)	0.964 (0.002)	0.910 (0.003)
SRHS >= fair	0.986 (0.001)	0.990 (0.001)	0.978 (0.001)

Note: For a given sampled parameter, z_t, this table reports an OLS estimate of ϕ_1 from the regression $z_t = \phi_0 + \phi_1 z_{t-1} + e_t$ with its Newey-West standard error in parentheses from the final 19,000 iterations.

Table 5: Sibling Correlations: Core Results

Outcome	Mean	Median	Std Dev
Asthma	0.486	0.486	0.031
Diabetes	0.628	0.630	0.085
Anemia	0.750	0.750	0.056
Development Delay	0.504	0.504	0.037
Hyperactivity	0.634	0.634	0.042
Allergies	0.569	0.568	0.045
Limitations on Athletics	0.532	0.532	0.042
Limitations on School Att.	0.633	0.630	0.054
Limitations on School Work	0.570	0.570	0.043
SRHS >= very good	0.628	0.628	0.040
SRHS >= good	0.614	0.613	0.042
SRHS >= fair	0.623	0.623	0.083
π^1	0.531	0.531	0.029
π^2	0.515	0.515	0.023
π^3	0.537	0.537	0.022
π^4	0.551	0.551	0.021
π^5	0.560	0.560	0.021
π^6	0.565	0.565	0.020

Note: We ran the sampler for 20,000 iterations but report results for the last 19,000 iterations.

Table 6: Sibling Correlations: Core Results by Gender

	Boys			Girls		
Outcome	Mean	Median	Std Dev	Mean	Median	Std Dev
Asthma	0.535	0.534	0.060	0.422	0.421	0.053
Diabetes	0.500	0.488	0.162	0.640	0.652	0.149
Anemia	0.716	0.730	0.114	0.756	0.765	0.096
Development Delay	0.717	0.720	0.073	0.528	0.525	0.084
Hyperactivity	0.730	0.733	0.077	0.576	0.572	0.087
Allergies	0.694	0.695	0.082	0.547	0.545	0.073
Limitations on Athletics	0.633	0.630	0.085	0.557	0.553	0.081
Limitations on School Att.	0.616	0.612	0.095	0.660	0.659	0.102
Limitations on School Work	0.630	0.628	0.082	0.665	0.667	0.084
SRHS >= very good	0.662	0.664	0.072	0.678	0.679	0.069
SRHS >= good	0.771	0.778	0.094	0.681	0.684	0.069
SRHS >= fair	0.821	0.843	0.113	0.645	0.657	0.173
π^1	0.624	0.626	0.054	0.540	0.540	0.049
π^2	0.603	0.604	0.043	0.523	0.523	0.041
π^3	0.625	0.627	0.040	0.548	0.549	0.040
π^4	0.642	0.644	0.039	0.557	0.557	0.039
π^5	0.649	0.650	0.038	0.561	0.561	0.038
π^6	0.653	0.655	0.038	0.561	0.562	0.038

Note: We ran the sampler for 20,000 iterations but report results for the last 19,000 iterations.

Table 7: Sibling Correlations: REML Estimates

Outcome	Estimate (Std Err)	95% Confidence Interval
Asthma	0.277 (0.040)	$[0.199, 0.354]$
Diabetes	0.209 (0.072)	$[0.068, 0.351]$
Anemia	0.926 (0.111)	$[0.708, 1.143]$
Development Delay	0.308 (0.058)	$[0.194, 0.421]$
Hyperactivity	0.314 (0.048)	$[0.220, 0.408]$
Allergies	0.725 (0.083)	$[0.562, 0.889]$
Limitations on Athletics	0.430 (0.075)	$[0.283, 0.576]$
Limitations on School Att.	0.827 (0.240)	$[0.358, 1.300]$
Limitations on School Work	0.165 (0.076)	$[0.016, 0.315]$
SRHS >= very good	0.916 (0.060)	$[0.798, 1.034]$
SRHS >= good	0.930 (0.069)	$[0.794, 1.066]$
SRHS >= fair	0.926 (0.012)	$[0.903, 0.949]$

Note: Estimates of ρ^m from model (11).

Table 8: REML Monte Carlo, m = 1

	$\sigma^2_\alpha = 2$		$\sigma^2_\gamma = 1$		$\rho = 0.667$	
Sample	LV	M	LV	M	LV	M
1	2.095	0.030	0.967	0.022	0.684	0.568
	(0.089)	(0.002)	(0.041)	(0.001)	(0.014)	(0.024)
2	2.073	0.029	1.012	0.023	0.672	0.564
	(0.089)	(0.002)	(0.043)	(0.001)	(0.015)	(0.025)
3	2.027	0.029	0.990	0.021	0.672	0.576
	(0.088)	(0.002)	(0.043)	(0.001)	(0.015)	(0.025)
4	2.044	0.034	0.992	0.022	0.673	0.604
	(0.089)	(0.002)	(0.042)	(0.001)	(0.015)	(0.023)
5	2.088	0.032	0.917	0.019	0.695	0.623
	(0.088)	(0.002)	(0.040)	(0.001)	(0.014)	(0.024)

Notes: LV corresponds to Monte Carlo simulations in which REML was estimated using the latent variable and M corresponds to simulations in which REML was estimated using the measurements as in equation (11). Standard errors in parentheses.

Table 9: REML Monte Carlo, m=2

	$\sigma^2_\alpha = 2$		$\sigma^2_\gamma = 1.5$		$\rho = 0.571$	
Sample	LV	M	LV	M	LV	M
1	2.051	0.018	1.591	0.025	0.563	0.426
	(0.010)	(0.001)	(0.061)	(0.001)	(0.017)	(0.026)
2	1.952	0.016	1.485	0.022	0.568	0.419
	(0.095)	(0.001)	(0.057)	(0.001)	(0.017)	(0.028)
3	2.011	0.017	1.350	0.020	0.598	0.458
	(0.094)	(0.001)	(0.053)	(0.001)	(0.017)	(0.027)
4	1.958	0.019	1.527	0.026	0.562	0.428
	(0.096)	(0.001)	(0.059)	(0.001)	(0.017)	(0.026)
5	1.867	0.015	1.511	0.022	0.553	0.410
	(0.093)	(0.001)	(0.058)	(0.001)	(0.018)	(0.028)

Notes: Per Table 8

Figure 1: CUMSUM Statistics

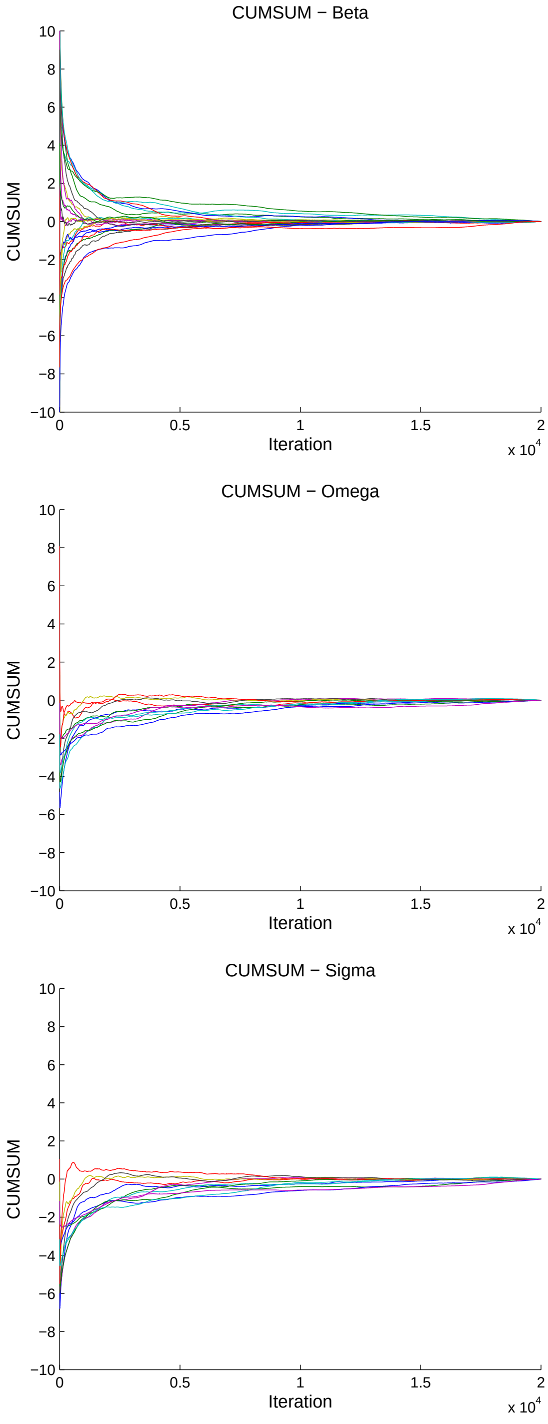


Figure 2: Time Series for Family and Individual Variances

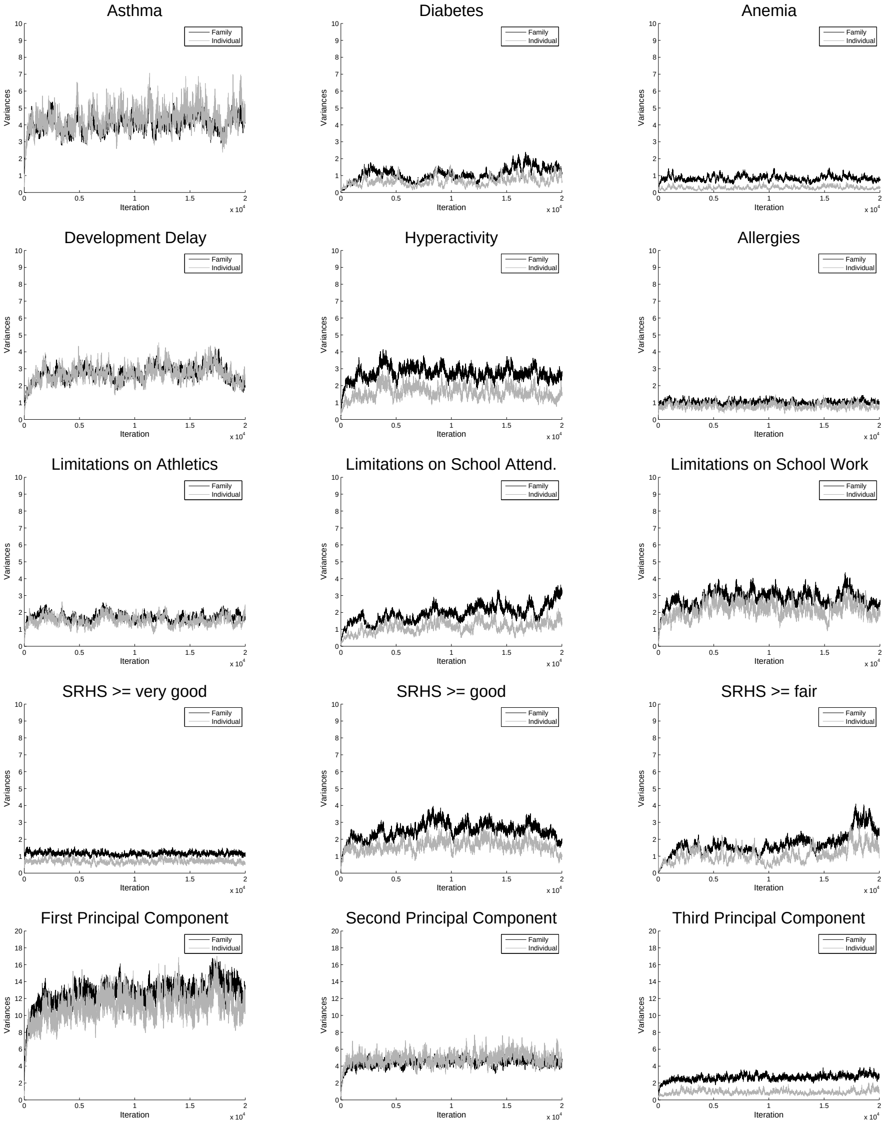
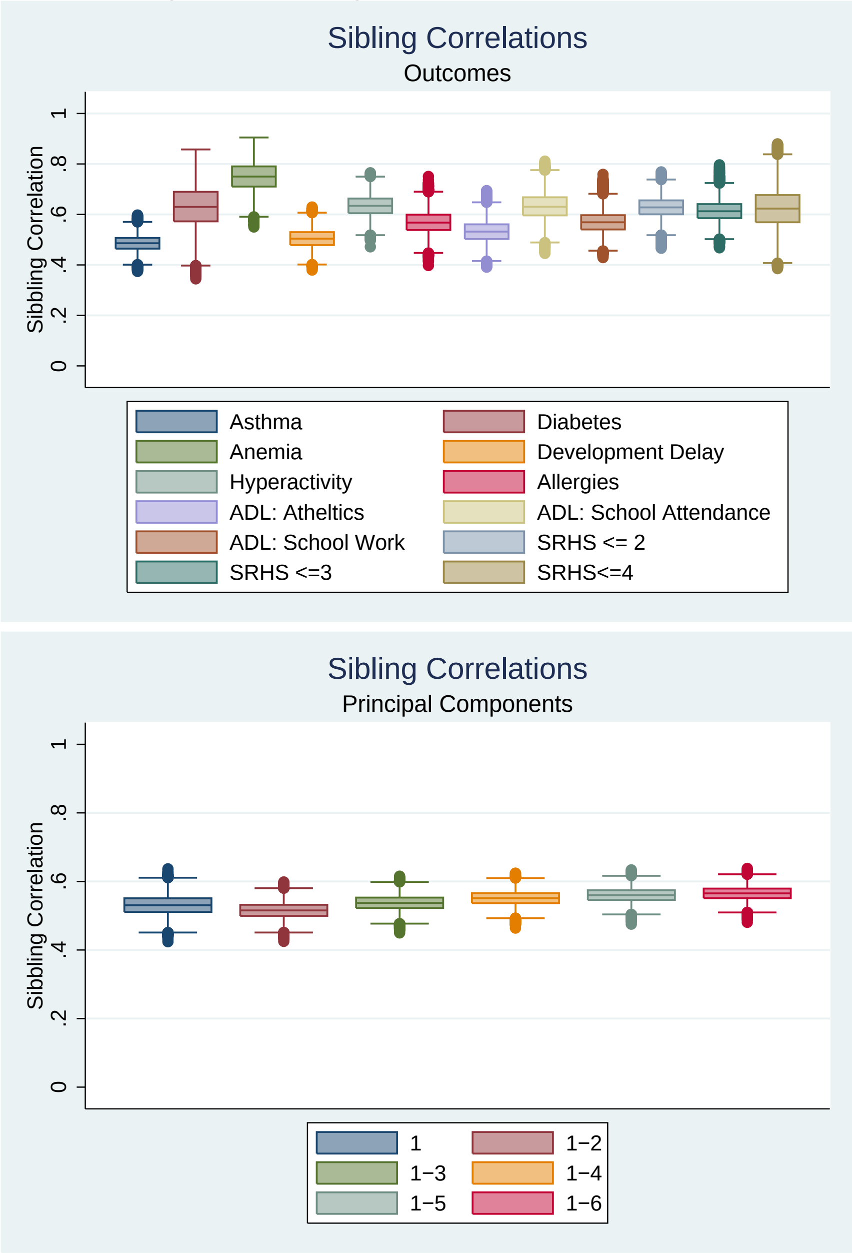


Figure 3: Sibling Correlation Distributions



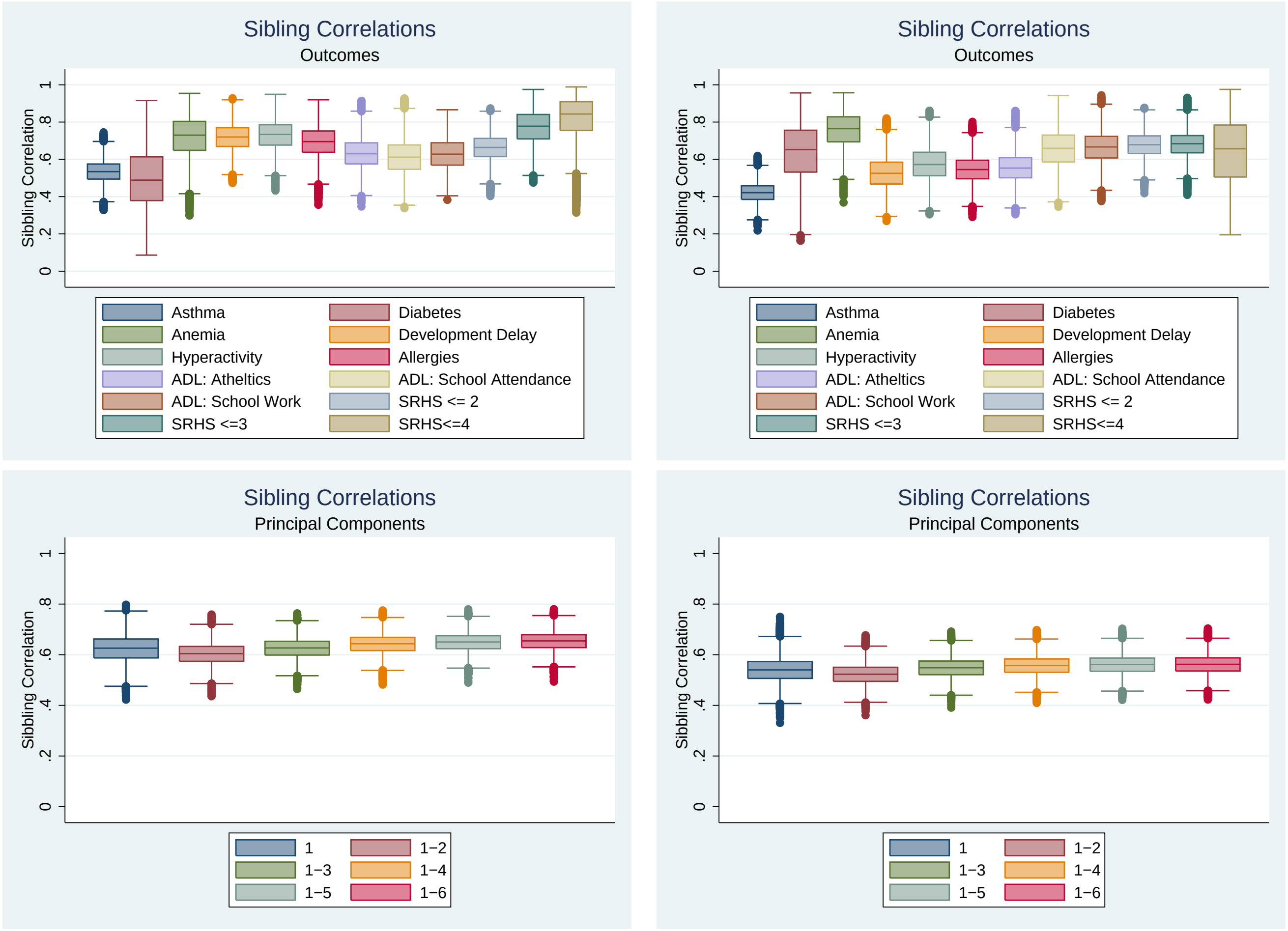
Sibling Correlations

Outcomes

Sibling Correlations

Principal Components

Figure 4: Sibling Correlation Distributions - Boys and Girls



Boys

Girls