Estimating Sibling Spillover Effects with Unobserved Confounding Using Gain-Scores

Appendix

David C. Mallinson & Felix Elwert

Contents

1 Overview

Consider linear data generating models, represented by **Figures 1-3** in the main text, with exposures T_{ij}, outcomes Y_{ij}, and unobserved confounders U_i for $i = 1, ..., N$ families consisting of $j = 1, 2$ siblings. Write $D_i = Y_{i2} - Y_{i1}$ for the gain-score. Henceforth, we suppress i for readability.

To estimate spillover effects from the exposure of one sibling onto the other sibling's outcome in these models, we use gain-score regression,

$$D = b_1 T_1 + b_2 T_2 + e,$$

where b_1 and b_2 are partial regression coefficients for T_1 and T_2, respectively, e is the residual, and $SC = b_1 + b_2$ is the spillover coefficient. The partial regression coefficients are given by,

$$b_1 = \frac{\sigma_{DT_1} - \sigma_{DT_2} * \sigma_{T_1T_2}}{1 - \sigma_{T_1T_2}{}^2},$$

$$b_2 = \frac{\sigma_{DT_2} - \sigma_{DT_1} * \sigma_{T_1T_2}}{1 - \sigma_{T_1T_2}{}^2},$$

where σ denotes covariance. We link the partial regression coefficients to the data generating models of **Figures 1-3** using Wright's path rules, where the covariance between two variables is equal to the sum of the products of path parameters on the open paths between the two variables. We assume standardized variables (mean zero and unit variance) without loss of generality. The identification results in the following sections hold for both continuous and binary exposures. See Pearl (2013) for details [1].

2 Models with One-Sided Spillover

Base model with one-sided spillover (Figure 1A)

We first consider models with one-sided exposure-to-outcome spillover ($T_1 \to Y_2$, but T_2 does not directly affect Y_1). Under the data generating process in **Figure 1A**, we point identify the spillover effect θ. We compute σ_{DT_1}, σ_{DT_2}, and $\sigma_{T_1T_2}$:

$$\begin{aligned}\sigma_{DT_1} &= -\psi\chi - \delta + \theta + \delta\chi\gamma + \psi\chi \\ &= -\delta + \theta + \delta\chi\gamma\end{aligned}$$

$$\begin{aligned}\sigma_{DT_2} &= -\psi\gamma - \delta\chi\gamma + \theta\chi\gamma + \delta + \psi\gamma \\ &= -\delta\chi\gamma + \theta\chi\gamma + \delta\end{aligned}$$

$$\sigma_{T_1T_2} = \chi\gamma$$

We plug these covariances into the formulas for our partial regression coefficients:

$$\begin{aligned}b_1 &= \frac{-\delta + \theta + \delta\chi\gamma - (-\delta\chi\gamma + \theta\chi\gamma + \delta) * (\chi\gamma)}{1 - (\chi\gamma)^2} \\ &= \frac{-\delta + \theta + \delta\chi\gamma - (-\delta\chi^2\gamma^2 + \theta\chi^2\gamma^2 + \delta\chi\gamma)}{1 - \chi^2\gamma^2} \\ &= \frac{-\delta + \theta + \delta\chi^2\gamma^2 - \theta\chi^2\gamma^2}{1 - \chi^2\gamma^2} \\ &= \frac{-\delta(1 - \chi^2\gamma^2) + \theta(1 - \chi^2\gamma^2)}{1 - \chi^2\gamma^2} \\ &= \theta - \delta\end{aligned}$$

$$\begin{aligned}b_2 &= \frac{-\delta\chi\gamma + \theta\chi\gamma + \delta - (-\delta + \theta + \delta\chi\gamma) * (\chi\gamma)}{1 - (\chi\gamma)^2} \\ &= \frac{-\delta\chi\gamma + \theta\chi\gamma + \delta - (-\delta\chi\gamma + \theta\chi\gamma + \delta\chi^2\gamma^2)}{1 - \chi^2\gamma^2} \\ &= \frac{\delta - \delta\chi^2\gamma^2}{1 - \chi^2\gamma^2} \\ &= \frac{\delta(1 - \chi^2\gamma^2)}{1 - \chi^2\gamma^2} \\ &= \delta\end{aligned}$$

We then compute SC, which point identifies θ:

$$\begin{aligned} SC &= b_1 + b_2 \\ &= \theta - \delta + \delta \\ &= \theta \end{aligned}$$

One-sided spillover and exposure-to-exposure spillover (Figures 1B-1C)

Under the data generating process in **Figure 1B**, sibling 2's exposure affects sibling 1's exposure ($T_2 \rightarrow T_1$). Regardless, we still point identify the spillover effect θ. We compute σ_{DT_1}, σ_{DT_2}, and $\sigma_{T_1T_2}$:

$$\begin{aligned} \sigma_{DT_1} &= -\psi\chi - \psi\gamma\tau - \delta + \theta + \delta\tau + \delta\chi\gamma + \psi\gamma\tau + \psi\chi \\ &= -\delta + \theta + \delta\tau + \delta\chi\gamma \end{aligned}$$

$$\begin{aligned} \sigma_{DT_2} &= -\psi\gamma - \delta\chi\gamma - \delta\tau + \theta\chi\gamma + \theta\tau + \delta + \psi\gamma \\ &= -\delta\chi\gamma - \delta\tau + \theta\chi\gamma + \theta\tau + \delta \end{aligned}$$

$$\sigma_{T_1T_2} = \chi\gamma + \tau$$

We then compute b_1 and b_2:

$$\begin{aligned} b_1 &= \frac{-\delta + \theta + \delta\tau + \delta\chi\gamma - (-\delta\chi\gamma - \delta\tau + \theta\chi\gamma + \theta\tau + \delta) * (\chi\gamma + \tau)}{1 - (\chi\gamma + \tau)^2} \\ &= \theta - \delta \end{aligned}$$

$$\begin{aligned} b_2 &= \frac{-\delta\chi\gamma - \delta\tau + \theta\chi\gamma + \theta\tau + \delta - (-\delta + \theta + \delta\tau + \delta\chi\gamma) * (\chi\gamma + \tau)}{1 - (\chi\gamma + \tau)^2} \\ &= \delta \end{aligned}$$

Finally, we compute SC:

$$\begin{aligned} SC &= b_1 + b_2 \\ &= \theta - \delta + \delta \\ &= \theta \end{aligned}$$

The data generating process in **Figure 1C** has exposure-to-exposure spillover, where sibling 1's exposure affects sibling 2's exposure ($T_1 \rightarrow T_2$). Still, we point identify the spillover effect θ. We compute σ_{DT_1}, σ_{DT_2}, and $\sigma_{T_1T_2}$:

$$\begin{aligned}\sigma_{DT_1} &= -\psi\chi - \delta + \theta + \delta\phi + \delta\chi\gamma + \psi\chi \\ &= -\delta + \theta + \delta\phi + \delta\chi\gamma\end{aligned}$$

$$\begin{aligned}\sigma_{DT_2} &= -\psi\gamma - \psi\chi\phi - \delta\chi\gamma - \delta\phi + \theta\chi\gamma + \theta\phi + \delta + \psi\chi\phi + \psi\gamma \\ &= -\delta\chi\gamma - \delta\phi + \theta\chi\gamma + \theta\phi + \delta\end{aligned}$$

$$\sigma_{T_1T_2} = \chi\gamma + \phi$$

We then compute b_1 and b_2:

$$\begin{aligned}b_1 &= \frac{-\delta + \theta + \delta\phi + \delta\chi\gamma - (-\delta\chi\gamma - \delta\phi + \theta\chi\gamma + \theta\phi + \delta) * (\chi\gamma + \phi)}{1 - (\chi\gamma + \phi)^2} \\ &= \theta - \delta\end{aligned}$$

$$\begin{aligned}b_2 &= \frac{-\delta\chi\gamma - \delta\phi + \theta\chi\gamma + \theta\phi + \delta - (-\delta + \theta + \delta\phi + \delta\chi\gamma) * (\chi\gamma + \phi)}{1 - (\chi\gamma + \phi)^2} \\ &= \delta\end{aligned}$$

Finally, we compute SC:

$$\begin{aligned}SC &= b_1 + b_2 \\ &= \theta - \delta + \delta \\ &= \theta\end{aligned}$$

3 Models with Two-Sided Spillover

Base model with two-sided spillover (Figure 2A)

We now consider models with two-sided exposure-to-outcome spillover ($T_1 \rightarrow Y_2$ and $T_2 \rightarrow Y_1$). Under the data generating process in **Figure 2A**, we identify the difference between θ and κ. We compute σ_{DT_1}, σ_{DT_2}, and $\sigma_{T_1T_2}$:

$$
\begin{aligned}
\sigma_{DT_1} &= -\psi\chi - \delta - \kappa\chi\gamma + \theta + \delta\chi\gamma + \psi\chi \\
&= -\delta - \kappa\chi\gamma + \theta + \delta\chi\gamma
\end{aligned}
$$

$$
\begin{aligned}
\sigma_{DT_2} &= -\psi\gamma - \delta\chi\gamma - \kappa + \theta\chi\gamma + \delta + \psi\gamma \\
&= -\delta\chi\gamma - \kappa + \theta\chi\gamma + \delta
\end{aligned}
$$

$$
\sigma_{T_1T_2} = \chi\gamma
$$

We then compute b_1 and b_2:

$$
\begin{aligned}
b_1 &= \frac{-\delta - \kappa\chi\gamma + \theta + \delta\chi\gamma - (-\delta\chi\gamma - \kappa + \theta\chi\gamma + \delta) * (\chi\gamma)}{1 - (\chi\gamma)^2} \\
&= \theta - \delta
\end{aligned}
$$

$$
\begin{aligned}
b_2 &= \frac{-\delta\chi\gamma - \kappa + \theta\chi\gamma + \delta - (-\delta - \kappa\chi\gamma + \theta + \delta\chi\gamma) * (\chi\gamma)}{1 - (\chi\gamma)^2} \\
&= \delta - \kappa
\end{aligned}
$$

Finally, we compute SC:

$$
\begin{aligned}
SC &= b_1 + b_2 \\
&= \theta - \delta + \delta - \kappa \\
&= \theta - \kappa
\end{aligned}
$$

Two-sided spillover and exposure-to-exposure spillover (Figures 2B-2C)

Even with exposure-to-exposure spillover, the spillover coefficient continues to identify the difference between the two exposure-to-outcome spillover effects ($SC = \theta - \kappa$). This is true regardless of the direction of the exposure-to-exposure spillover. In **Figure 2B**, we have exposure-to-exposure spillover from sibling 2 to sibling 1 ($T_2 \to T_1$). We compute σ_{DT_1}, σ_{DT_2}, and $\sigma_{T_1T_2}$:

$$\begin{aligned}\sigma_{DT_1} &= -\psi\chi - \psi\gamma\tau - \delta - \kappa\tau - \kappa\chi\gamma + \theta + \delta\tau + \delta\chi\gamma + \psi\gamma\tau + \psi\chi \\ &= -\delta - \kappa\tau - \kappa\chi\gamma + \theta + \delta\tau + \delta\chi\gamma\end{aligned}$$

$$\begin{aligned}\sigma_{DT_2} &= -\psi\gamma - \delta\chi\gamma + \delta\tau - \kappa + \theta\chi\gamma + \theta\tau + \delta + \psi\gamma \\ &= -\delta\chi\gamma + \delta\tau - \kappa + \theta\chi\gamma + \theta\tau + \delta\end{aligned}$$

$$\sigma_{T_1T_2} = \chi\gamma + \tau$$

We then compute b_1 and b_2:

$$\begin{aligned}b_1 &= \frac{-\delta - \kappa\tau - \kappa\chi\gamma + \theta + \delta\tau + \delta\chi\gamma - (-\delta\chi\gamma + \delta\tau - \kappa + \theta\chi\gamma + \theta\tau + \delta) * (\chi\gamma + \tau)}{1 - (\chi\gamma + \tau)^2} \\ &= \theta - \delta\end{aligned}$$

$$\begin{aligned}b_2 &= \frac{-\delta\chi\gamma + \delta\tau - \kappa + \theta\chi\gamma + \theta\tau + \delta - (-\delta - \kappa\tau - \kappa\chi\gamma + \theta + \delta\tau + \delta\chi\gamma) * (\chi\gamma + \tau)}{1 - (\chi\gamma + \tau)^2} \\ &= \delta - \kappa\end{aligned}$$

Finally, we compute SC:

$$\begin{aligned}SC &= b_1 + b_2 \\ &= \theta - \delta + \delta - \kappa \\ &= \theta - \kappa\end{aligned}$$

In **Figure 2C**, we have exposure-to-exposure spillover from sibling 1 to sibling 2 ($T_1 \to T_2$). Changing the direction of the exposure-to-exposure spillover does not affect the partial regression coefficients nor the spillover coefficient.

4 Models with Spillovers from Outcomes

Past outcome affects future exposure (Figure 3A)

We consider models in which spillovers originate from outcomes. Under the data generating process in **Figure 3A**, we do not identify the spillover effect θ. We compute σ_{DT_1}, σ_{DT_2}, and $\sigma_{T_1T_2}$:

$$\begin{aligned}\sigma_{DT_1} &= -\psi\chi - \delta + \theta + \delta\chi\gamma + \delta^2\omega + \delta\omega\psi\chi + \psi\chi \\ &= -\delta + \theta + \delta\chi\gamma + \delta^2\omega + \delta\omega\psi\chi\end{aligned}$$

$$\begin{aligned}\sigma_{DT_2} &= -\psi\gamma - \psi\chi\delta\omega - \delta\chi\gamma - \omega + \theta\chi\gamma + \theta\delta\omega + \delta + \psi^2\omega + \psi\chi\delta\omega + \psi\gamma \\ &= -\delta\chi\gamma - \omega + \theta\chi\gamma + \theta\delta\omega + \delta + \psi^2\omega\end{aligned}$$

$$\sigma_{T_1T_2} = \chi\gamma + \delta\omega + \omega\psi\chi$$

We then compute b_1 and b_2:

$$b_1 = \frac{-\delta + \theta + \delta\chi\gamma + \delta^2\omega + \delta\omega\psi\chi - (-\delta\chi\gamma - \omega + \theta\chi\gamma + \theta\delta\omega + \delta + \psi^2\omega) * (\chi\gamma + \delta\omega + \omega\psi\chi)}{1 - (\chi\gamma + \delta\omega + \omega\psi\chi)^2}$$

$$b_2 = \frac{-\delta\chi\gamma - \omega + \theta\chi\gamma + \theta\delta\omega + \delta + \psi^2\omega - (-\delta + \theta + \delta\chi\gamma + \delta^2\omega + \delta\omega\psi\chi) * (\chi\gamma + \delta\omega + \omega\psi\chi)}{1 - (\chi\gamma + \delta\omega + \omega\psi\chi)^2}$$

Finally, we compute SC:

$$\begin{aligned}SC &= b_1 + b_2 \\ &= \frac{-\delta + \theta + \delta\chi\gamma + \delta^2\omega + \delta\omega\psi\chi - (-\delta\chi\gamma - \omega + \theta\chi\gamma + \theta\delta\omega + \delta + \psi^2\omega) * (\chi\gamma + \delta\omega + \omega\psi\chi)}{1 - (\chi\gamma + \delta\omega + \omega\psi\chi)^2} + \\ &\quad \frac{-\delta\chi\gamma - \omega + \theta\chi\gamma + \theta\delta\omega + \delta + \psi^2\omega - (-\delta + \theta + \delta\chi\gamma + \delta^2\omega + \delta\omega\psi\chi) * (\chi\gamma + \delta\omega + \omega\psi\chi)}{1 - (\chi\gamma + \delta\omega + \omega\psi\chi)^2} \\ &= \frac{\begin{array}{l}\theta + \delta^2\omega + \delta\omega\psi\chi - \omega + \theta\chi\gamma + \theta\delta\omega + \psi^2\omega - \\ \quad (-\omega + \theta\chi\gamma + \theta\delta\omega + \psi^2\omega + \theta + \delta^2\omega + \delta\omega\psi\chi) * (\chi\gamma + \delta\omega + \omega\psi\chi)\end{array}}{1 - (\chi\gamma + \delta\omega + \omega\psi\chi)^2}\end{aligned}$$

In this data generating process, computing $SC = b_1 + b_2$ does not identify θ nor a simple function of θ.

Outcome-to-outcome spillover (Figures 3B-3C)

Under the data generating process in **Figure 3B**, sibling 1's outcome affects sibling 2's outcome ($Y_1 \rightarrow Y_2$), and we do not identify the spillover effect θ. We compute σ_{DT_1}, σ_{DT_2}, and $\sigma_{T_1T_2}$:

$$\begin{aligned}\sigma_{DT_1} &= -\psi\chi - \delta + \eta\psi\gamma + \eta\delta + \theta + \delta\chi\gamma + \psi\chi \\ &= -\delta + \eta\psi\gamma + \eta\delta + \theta + \delta\chi\gamma\end{aligned}$$

$$\begin{aligned}\sigma_{DT_2} &= -\psi\gamma - \delta\chi\gamma + \eta\psi\gamma + \eta\delta\chi\gamma + \theta\chi\gamma + \delta + \psi\gamma \\ &= -\delta\chi\gamma + \eta\psi\gamma + \eta\delta\chi\gamma + \theta\chi\gamma + \delta\end{aligned}$$

$$\sigma_{T_1T_2} = \chi\gamma$$

We then compute b_1 and b_2:

$$\begin{aligned}b_1 &= \frac{-\delta + \eta\psi\gamma + \eta\delta + \theta + \delta\chi\gamma - (-\delta\chi\gamma + \eta\psi\gamma + \eta\delta\chi\gamma + \theta\chi\gamma + \delta) * (\chi\gamma)}{1 - (\chi\gamma)^2} \\ &= -\delta + \theta + \eta\delta + \frac{\eta\psi\gamma - \eta\psi\chi\gamma^2}{1 - \chi^2\gamma^2}\end{aligned}$$

$$\begin{aligned}b_2 &= \frac{-\delta\chi\gamma + \eta\psi\gamma + \eta\delta\chi\gamma + \theta\chi\gamma + \delta - (-\delta + \eta\psi\gamma + \eta\delta + \theta + \delta\chi\gamma) * (\chi\gamma)}{1 - (\chi\gamma)^2} \\ &= \delta + \frac{\eta\psi\gamma - \eta\psi\chi\gamma^2}{1 - \chi^2\gamma^2}\end{aligned}$$

Finally, we compute SC:

$$\begin{aligned}SC &= b_1 + b_2 \\ &= -\delta + \theta + \eta\delta + \frac{\eta\psi\gamma - \eta\psi\chi\gamma^2}{1 - \chi^2\gamma^2} + \delta + \frac{\eta\psi\gamma - \eta\psi\chi\gamma^2}{1 - \chi^2\gamma^2} \\ &= \theta + \eta\delta + \frac{2\eta\psi\gamma - 2\eta\psi\chi\gamma^2}{1 - \chi^2\gamma^2}\end{aligned}$$

In this data generating process, computing $SC = b_1 + b_2$ does not identify θ nor a simple function of θ.

The data generating process in **Figure 3C** has outcome-to-outcome spillover in which sibling 2's outcome affects sibling 1's outcome ($Y_2 \rightarrow Y_1$). Similarly, we do not identify the spillover effect θ. We compute σ_{DT_1}, σ_{DT_2}, and $\sigma_{T_1T_2}$:

$$\begin{aligned}\sigma_{DT_1} &= -\psi\chi - \delta - \lambda\theta - \lambda\delta\chi\gamma - \lambda\psi\chi + \theta + \delta\chi\gamma + \psi\chi \\ &= -\delta - \lambda\theta - \lambda\delta\chi\gamma - \lambda\psi\chi + \theta + \delta\chi\gamma\end{aligned}$$

$$
\begin{aligned}
\sigma_{DT_2} &= -\psi\gamma - \delta\chi\gamma - \lambda\theta\chi\gamma - \lambda\delta + \theta\chi\gamma + \delta + \psi\gamma \\
&= -\delta\chi\gamma - \lambda\theta\chi\gamma - \lambda\delta + \theta\chi\gamma + \delta
\end{aligned}
$$

$$
\sigma_{T_1T_2} = \chi\gamma
$$

We then compute b_1 and b_2:

$$
\begin{aligned}
b_1 &= \frac{-\delta - \lambda\theta - \lambda\delta\chi\gamma - \lambda\psi\chi + \theta + \delta\chi\gamma - (-\delta\chi\gamma - \lambda\theta\chi\gamma - \lambda\delta + \theta\chi\gamma + \delta) * (\chi\gamma)}{1 - (\chi\gamma)^2} \\
&= -\delta - \lambda\theta + \theta - \frac{\lambda\psi\chi}{1 - \chi^2\gamma^2}
\end{aligned}
$$

$$
\begin{aligned}
b_2 &= \frac{-\delta\chi\gamma - \lambda\theta\chi\gamma - \lambda\delta + \theta\chi\gamma + \delta - (-\delta - \lambda\theta - \lambda\delta\chi\gamma - \lambda\psi\chi + \theta + \delta\chi\gamma) * (\chi\gamma)}{1 - (\chi\gamma)^2} \\
&= \delta + \lambda\delta + \frac{\lambda\psi\chi^2\gamma}{1 - \chi^2\gamma^2}
\end{aligned}
$$

Finally, we compute SC:

$$
\begin{aligned}
SC &= b_1 + b_2 \\
&= -\delta - \lambda\theta + \theta - \frac{\lambda\psi\chi}{1 - \chi^2\gamma^2} + \delta + \lambda\delta + \frac{\lambda\psi\chi^2\gamma}{1 - \chi^2\gamma^2} \\
&= \theta - \lambda\theta + \frac{\lambda\psi\chi^2\gamma - \lambda\psi\chi}{1 - \chi^2\gamma^2}
\end{aligned}
$$

In this data generating process, computing $SC = b_1 + b_2$ does not identify θ nor a simple function of θ.

5 Data Assembly for the Empirical Application

For our application, we estimated the effect of a younger sibling's preterm birth (gestational age <37 completed weeks) on their older sibling's Phonological Awareness Literacy Screening-Kindergarten (PALS-K) test score. PALS-K evaluates fundamental literacy skills at kindergarten entry [2]. We used data from Big Data for Little Kids (BD4LK), a cohort of birth records for live resident in-state deliveries in Wisconsin during 2007-2016 (N>666,000 births). Birth records link to multiple administrative sources, including paid Medicaid claims and encounters (2007-2016) and children's PALS-K scores (2012-2016 school years). BD4LK's linkage process has been previously described [3, 4].

We sampled sibling pairs born sequentially to the same mother between January 1, 2007-September 1, 2010 and took the English-language PALS-K test. We restricted eligibility on birthdate because children had to be at least five years old by September 1, 2015 for PALS-K testing eligibility [5]. PALS-K is available in two languages: English and Spanish [2, 6]. However, test versions were developed separately and are not directly comparable, so we only considered children who took the English-language test.

BD4LK includes 252,883 unique deliveries during January 1, 2007-September 1, 2010. We identified 806 records (0.3%) with multiple maternal or child identifiers that imperfectly matched to Medicaid claims or PALS-K scores. Among those records, we randomly selected one match and excluded the remainder. We then identified 177,863 records (70.3%) that linked to PALS-K scores, of which 46,743 records were in-sample siblings (26.3% of test-linked records). Finally, we pulled eligible sibling pairs: sequentially-born from different deliveries, took the English-language PALS-K test, and completed information on control variables (maternal age, maternal education, and Medicaid delivery coverage, all of which were measured at the older sibling's delivery). This generated a final sample of 40,020 siblings (85.6% of tested siblings), or 20,010 sibling pairs.

6 Appendix Tables

Table A.1. Descriptive statistics of sibling pairs for the empirical application (N = 20,010 sibling pairs[a])

	Older Siblings (N = 20,010)	**Younger Siblings (N = 20,010)**
PALS-K Score (Points)[c], Mean (SD)	63.58 (24.12)	64.22 (23.83)
Preterm Birth[b], N (%)	1,357 (6.78%)	1,331 (6.65%)
Maternal Age (Years)[d], Mean (SD)	26.01 (5.12)	–
Maternal Education[d], N (%)		
No high school degree	2,865 (14.32%)	–
High school degree/equivalent only	5,789 (28.93%)	–
1-3 years college	5,036 (25.17%)	–
4+ years college	6,320 (31.58%)	–
Medicaid-Paid Delivery[d], N (%)	7,528 (37.62%)	–

[a]Each sibling pair includes one older sibling and one younger sibling. The sample consists of 40,020 children in total.

[b]Preterm birth is defined as gestational age <37 completed weeks.

[c]PALS-K has a score range of 0-102 points.

[d]Measured at the older sibling's delivery.

Notes: "PALS-K" Phonological Awareness Literacy Screening-Kindergarten, "SD" standard deviation.

Table A.2. Cross-tabulation of preterm birth[a] by sibling within two-sibling clusters (N = 20,010 pairs[b])

	Younger Sibling Not Preterm	**Younger Sibling Preterm**	**Total**
Older Sibling Not Preterm	17,652	1,001	18,653
	(94.63%)	(5.37%)	(100.00%)
	(94.50%)	(74.21%)	(93.22%)
Older Sibling Preterm	1,027	330	1,357
	(75.68%)	(24.32%)	(100.00%)
	(5.50%)	(24.79%)	(6.78%)
Total	18,679	1,331	20,010
	(93.35%)	(6.65%)	(100.00%)
	(100.00%)	(100.00%)	(100.00%)

[a]Preterm birth is defined as gestational age <37 completed weeks.

[b]Each sibling pair includes one older sibling and one younger sibling. The sample consists of 40,020 children in total.

Notes: Whole numbers indicate the frequency of sibling pairs. Within a cell, the first percentage indicates the row percentage (i.e., the percent of younger siblings who are preterm or not preterm by the older sibling's preterm birth status), and the second percentage is a column percentage (i.e., the percent of older siblings who are preterm or not preterm by the younger sibling's preterm birth status).

7 Simulation Code

We conducted simulations in Stata Statistical Software: Release 16 [7].

```
/*
Estimatinng Sibling Spillovers with
Unobserved Confounding Using Gain-Scores

Simulation Code

David C. Mallinson, Felix Elwert
*/

/*********************/
*Preliminaries
/*********************/

/*
Defintions:

t_1 = sibling 1's treatment
t_2 = sibling 2's treatment
y_1 = sibling 1's outcome
y_2 = sibling 2's outcome
u = unobserved fixed effect
d = gain-score (y_2-y_1)
*/

**set working directory

/*
NOTE: This simulation saves datasets to compile results across models. Saved
files will be in the working directory. To specify the directory, run the cd
command with the desired file path (without the *).

At the end of this file, there is code to delete all resulting files from
this simulation analysis. This code is optional (for cleanup) and must be run
manually. See end of code for details.
*/

*cd "[FILE PATH HERE]"


/*********************/
*Simulate Model 1A
/*********************/

**clear previous data
clear all
set more off
```

```
**monte carlo simulation settings
global nobs=5000 /*observations per run*/
global nmc=1000 /*number of runs*/
set seed 54385 /*seed for random number generator -- can change*/
set obs $nobs /*sets observations within each run*/

**set parameters (causal effects in models)
scalar delta=1 /*t_1 on y_1, t_2 on y_2*/

scalar psi=1 /*u on y_1, u on y_2*/
scalar chi=2 /*u on t_1*/
scalar gamma=3 /*u on t_2*/

scalar theta=0.5 /*t_1 on y_2*/
scalar kappa=0.3 /*t_2 on y_1*/
scalar tau=0.3 /*t_2 on t_1*/
scalar phi=0.3 /*t_1 on t_2*/

scalar omega=0.3 /*y_1 on t_2*/
scalar eta=0.3 /*y_1 on y_2*/
scalar lambda=0.3 /*y_2 on y_1*/

**generate variables

gen u=. /*unobserved fixed effect*/
gen t_1=. /*sibling 1's treatment*/
gen t_2=. /*sibling 2's treatment*/
gen y_1=. /*sibling 1's outcome*/
gen y_2=. /*sibling 2's outcome*/
gen d=. /*gain-score (y_2-y_1)*/

**simulation

tempname m1a
postfile `m1a' sc_mc1 sc_ub_mc1 sc_lb_mc1 using m1a, replace
quietly {

forvalues i=1/$nmc {

   /*note: for y_1 and y_2 equations, "rnormal(0,1)" is residual*/

   replace u=rnormal(0,1)
   replace t_1=(chi*u)>0.5
   replace t_2=(gamma*u)>0.2
   replace y_1=(delta*t_1)+(psi*u)+rnormal(0,1)
   replace y_2=(delta*t_2)+(theta*t_1)+(psi*u)+rnormal(0,1)
   replace d=y_2-y_1

   /*gain-score regression*/
```

```
    reg d t_1 t_2

    /*lincom to compute spillover coefficient*/
    lincom t_1+t_2

    /*store lincom output*/
    scalar sc_mc1=r(estimate) /*spillover coefficient*/
    scalar sc_ub_mc1=(r(estimate)+(1.96*r(se))) /*95% CI upper bound*/
    scalar sc_lb_mc1=(r(estimate)-(1.96*r(se))) /*95% CI lower bound*/

    post `m1a' (sc_mc1) (sc_ub_mc1) (sc_lb_mc1)
}
}
postclose `m1a'

use m1a, clear
summarize

**store mean of results, save to append with other simulation results
egen sc=mean(sc_mc1)
egen ub=mean(sc_ub_mc1)
egen lb=mean(sc_lb_mc1)
gen id=1 /*first simulation*/

keep sc ub lb id
duplicates drop
duplicates report

**save data
save "sim1.dta", replace


/*********************/
*Simulate Model 1B
/*********************/

**clear previous data
clear all
set more off

**monte carlo simulation settings
global nobs=5000 /*observations per run*/
global nmc=1000 /*number of runs*/
set seed 2414830 /*seed for random number generator -- can change*/
set obs $nobs /*sets observations within each run*/

**set parameters (causal effects in models)
scalar delta=1 /*t_1 on y_1, t_2 on y_2*/

scalar psi=1 /*u on y_1, u on y_2*/
```

```
scalar chi=2 /*u on t_1*/
scalar gamma=3 /*u on t_2*/

scalar theta=0.5 /*t_1 on y_2*/
scalar kappa=0.3 /*t_2 on y_1*/
scalar tau=0.3 /*t_2 on t_1*/
scalar phi=0.3 /*t_1 on t_2*/

scalar omega=0.3 /*y_1 on t_2*/
scalar eta=0.3 /*y_1 on y_2*/
scalar lambda=0.3 /*y_2 on y_1*/

**generate variables

gen u=. /*unobserved fixed effect*/
gen t_1=. /*sibling 1's treatment*/
gen t_2=. /*sibling 2's treatment*/
gen y_1=. /*sibling 1's outcome*/
gen y_2=. /*sibling 2's outcome*/
gen d=. /*gain-score (y_2-y_1)*/

**simulation

tempname m1b
postfile `m1b' sc_mc2 sc_ub_mc2 sc_lb_mc2 using m1b, replace
quietly {

forvalues i=1/$nmc {

   /*note: for y_1 and y_2 equations, "rnormal(0,1)" is residual*/

   replace u=rnormal(0,1)
   replace t_2=(gamma*u)>0.2
   replace t_1=(tau*t_2)+(chi*u)>0.5
   replace y_1=(delta*t_1)+(psi*u)+rnormal(0,1)
   replace y_2=(delta*t_2)+(theta*t_1)+(psi*u)+rnormal(0,1)
   replace d=y_2-y_1

   /*gain-score regression*/
   reg d t_1 t_2

   /*lincom to compute spillover coefficient*/
   lincom t_1+t_2

   /*store lincom output*/
   scalar sc_mc2=r(estimate) /*spillover coefficient*/
   scalar sc_ub_mc2=(r(estimate)+(1.96*r(se))) /*95% CI upper bound*/
   scalar sc_lb_mc2=(r(estimate)-(1.96*r(se))) /*95% CI lower bound*/

   post `m1b' (sc_mc2) (sc_ub_mc2) (sc_lb_mc2)
```

```
}
}
postclose `m1b'

use m1b, clear
summarize

**store mean of results, save to append with other simulation results
egen sc=mean(sc_mc2)
egen ub=mean(sc_ub_mc2)
egen lb=mean(sc_lb_mc2)
gen id=2 /*second simulation*/

keep sc ub lb id
duplicates drop
duplicates report

**save data
save "sim2.dta", replace


/*********************/
*Simulate Model 1C
/*********************/

**clear previous data
clear all
set more off

**monte carlo simulation settings
global nobs=5000 /*observations per run*/
global nmc=1000 /*number of runs*/
set seed 48910 /*seed for random number generator -- can change*/
set obs $nobs /*sets observations within each run*/

**set parameters (causal effects in models)
scalar delta=1 /*t_1 on y_1, t_2 on y_2*/

scalar psi=1 /*u on y_1, u on y_2*/
scalar chi=2 /*u on t_1*/
scalar gamma=3 /*u on t_2*/

scalar theta=0.5 /*t_1 on y_2*/
scalar kappa=0.3 /*t_2 on y_1*/
scalar tau=0.3 /*t_2 on t_1*/
scalar phi=0.3 /*t_1 on t_2*/

scalar omega=0.3 /*y_1 on t_2*/
scalar eta=0.3 /*y_1 on y_2*/
scalar lambda=0.3 /*y_2 on y_1*/
```

```
**generate variables

gen u=. /*unobserved fixed effect*/
gen t_1=. /*sibling 1's treatment*/
gen t_2=. /*sibling 2's treatment*/
gen y_1=. /*sibling 1's outcome*/
gen y_2=. /*sibling 2's outcome*/
gen d=. /*gain-score (y_2-y_1)*/

**simulation

tempname m1c
postfile `m1c' sc_mc3 sc_ub_mc3 sc_lb_mc3 using m1c, replace
quietly {

forvalues i=1/$nmc {

   /*note: for y_1 and y_2 equations, "rnormal(0,1)" is residual*/

   replace u=rnormal(0,1)
   replace t_1=(chi*u)>0.5
   replace t_2=(phi*t_1)+(gamma*u)>0.2
   replace y_1=(delta*t_1)+(psi*u)+rnormal(0,1)
   replace y_2=(delta*t_2)+(theta*t_1)+(psi*u)+rnormal(0,1)
   replace d=y_2-y_1

   /*gain-score regression*/
   reg d t_1 t_2

   /*lincom to compute spillover coefficient*/
   lincom t_1+t_2

   /*store lincom output*/
   scalar sc_mc3=r(estimate) /*spillover coefficient*/
   scalar sc_ub_mc3=(r(estimate)+(1.96*r(se))) /*95% CI upper bound*/
   scalar sc_lb_mc3=(r(estimate)-(1.96*r(se))) /*95% CI lower bound*/

   post `m1c' (sc_mc3) (sc_ub_mc3) (sc_lb_mc3)
}
}
postclose `m1c'

use m1c, clear
summarize

**store mean of results, save to append with other simulation results
egen sc=mean(sc_mc3)
egen ub=mean(sc_ub_mc3)
egen lb=mean(sc_lb_mc3)
```

```
gen id=3 /*third simulation*/

keep sc ub lb id
duplicates drop
duplicates report

**save data
save "sim3.dta", replace


/*********************/
*Simulate Model 2A
/*********************/

**clear previous data
clear all
set more off

**monte carlo simulation settings
global nobs=5000 /*observations per run*/
global nmc=1000 /*number of runs*/
set seed 9842 /*seed for random number generator -- can change*/
set obs $nobs /*sets observations within each run*/

**set parameters (causal effects in models)
scalar delta=1 /*t_1 on y_1, t_2 on y_2*/

scalar psi=1 /*u on y_1, u on y_2*/
scalar chi=2 /*u on t_1*/
scalar gamma=3 /*u on t_2*/

scalar theta=0.5 /*t_1 on y_2*/
scalar kappa=0.3 /*t_2 on y_1*/
scalar tau=0.3 /*t_2 on t_1*/
scalar phi=0.3 /*t_1 on t_2*/

scalar omega=0.3 /*y_1 on t_2*/
scalar eta=0.3 /*y_1 on y_2*/
scalar lambda=0.3 /*y_2 on y_1*/

**generate variables

gen u=. /*unobserved fixed effect*/
gen t_1=. /*sibling 1's treatment*/
gen t_2=. /*sibling 2's treatment*/
gen y_1=. /*sibling 1's outcome*/
gen y_2=. /*sibling 2's outcome*/
gen d=. /*gain-score (y_2-y_1)*/

**simulation
```

```
tempname m2a
postfile `m2a' sc_mc4 sc_ub_mc4 sc_lb_mc4 using m2a, replace
quietly {

forvalues i=1/$nmc {

    /*note: for y_1 and y_2 equations, "rnormal(0,1)" is residual*/

    replace u=rnormal(0,1)
    replace t_1=(chi*u)>0.5
    replace t_2=(gamma*u)>0.2
    replace y_1=(delta*t_1)+(kappa*t_2)+(psi*u)+rnormal(0,1)
    replace y_2=(delta*t_2)+(theta*t_1)+(psi*u)+rnormal(0,1)
    replace d=y_2-y_1

    /*gain-score regression*/
    reg d t_1 t_2

    /*lincom to compute spillover coefficient*/
    lincom t_1+t_2

    /*store lincom output*/
    scalar sc_mc4=r(estimate) /*spillover coefficient*/
    scalar sc_ub_mc4=(r(estimate)+(1.96*r(se))) /*95% CI upper bound*/
    scalar sc_lb_mc4=(r(estimate)-(1.96*r(se))) /*95% CI lower bound*/

    post `m2a' (sc_mc4) (sc_ub_mc4) (sc_lb_mc4)
}
}
postclose `m2a'

use m2a, clear
summarize

**store mean of results, save to append with other simulation results
egen sc=mean(sc_mc4)
egen ub=mean(sc_ub_mc4)
egen lb=mean(sc_lb_mc4)
gen id=4 /*fourth simulation*/

keep sc ub lb id
duplicates drop
duplicates report

**save data
save "sim4.dta", replace


/*********************/
```

```
*Simulate Model 2B
/*********************/

**clear previous data
clear all
set more off

**monte carlo simulation settings
global nobs=5000 /*observations per run*/
global nmc=1000 /*number of runs*/
set seed 33305 /*seed for random number generator -- can change*/
set obs $nobs /*sets observations within each run*/

**set parameters (causal effects in models)
scalar delta=1 /*t_1 on y_1, t_2 on y_2*/

scalar psi=1 /*u on y_1, u on y_2*/
scalar chi=2 /*u on t_1*/
scalar gamma=3 /*u on t_2*/

scalar theta=0.5 /*t_1 on y_2*/
scalar kappa=0.3 /*t_2 on y_1*/
scalar tau=0.3 /*t_2 on t_1*/
scalar phi=0.3 /*t_1 on t_2*/

scalar omega=0.3 /*y_1 on t_2*/
scalar eta=0.3 /*y_1 on y_2*/
scalar lambda=0.3 /*y_2 on y_1*/

**generate variables

gen u=. /*unobserved fixed effect*/
gen t_1=. /*sibling 1's treatment*/
gen t_2=. /*sibling 2's treatment*/
gen y_1=. /*sibling 1's outcome*/
gen y_2=. /*sibling 2's outcome*/
gen d=. /*gain-score (y_2-y_1)*/

**simulation

tempname m2b
postfile `m2b' sc_mc5 sc_ub_mc5 sc_lb_mc5 using m2b, replace
quietly {

forvalues i=1/$nmc {

   /*note: for y_1 and y_2 equations, "rnormal(0,1)" is residual*/

   replace u=rnormal(0,1)
   replace t_2=(gamma*u)>0.2
```

```
   replace t_1=(tau*t_2)+(chi*u)>0.5
   replace y_1=(delta*t_1)+(kappa*t_2)+(psi*u)+rnormal(0,1)
   replace y_2=(delta*t_2)+(theta*t_1)+(psi*u)+rnormal(0,1)
   replace d=y_2-y_1

   /*gain-score regression*/
   reg d t_1 t_2

   /*lincom to compute spillover coefficient*/
   lincom t_1+t_2

   /*store lincom output*/
   scalar sc_mc5=r(estimate) /*spillover coefficient*/
   scalar sc_ub_mc5=(r(estimate)+(1.96*r(se))) /*95% CI upper bound*/
   scalar sc_lb_mc5=(r(estimate)-(1.96*r(se))) /*95% CI lower bound*/

   post `m2b' (sc_mc5) (sc_ub_mc5) (sc_lb_mc5)
}
}
postclose `m2b'

use m2b, clear
summarize

**store mean of results, save to append with other simulation results
egen sc=mean(sc_mc5)
egen ub=mean(sc_ub_mc5)
egen lb=mean(sc_lb_mc5)
gen id=5 /*fifth simulation*/

keep sc ub lb id
duplicates drop
duplicates report

**save data
save "sim5.dta", replace


/*********************/
*Simulate Model 2C
/*********************/

**clear previous data
clear all
set more off

**monte carlo simulation settings
global nobs=5000 /*observations per run*/
global nmc=1000 /*number of runs*/
set seed 101678 /*seed for random number generator -- can change*/
```

```
set obs $nobs /*sets observations within each run*/

**set parameters (causal effects in models)
scalar delta=1 /*t_1 on y_1, t_2 on y_2*/

scalar psi=1 /*u on y_1, u on y_2*/
scalar chi=2 /*u on t_1*/
scalar gamma=3 /*u on t_2*/

scalar theta=0.5 /*t_1 on y_2*/
scalar kappa=0.3 /*t_2 on y_1*/
scalar tau=0.3 /*t_2 on t_1*/
scalar phi=0.3 /*t_1 on t_2*/

scalar omega=0.3 /*y_1 on t_2*/
scalar eta=0.3 /*y_1 on y_2*/
scalar lambda=0.3 /*y_2 on y_1*/

**generate variables

gen u=. /*unobserved fixed effect*/
gen t_1=. /*sibling 1's treatment*/
gen t_2=. /*sibling 2's treatment*/
gen y_1=. /*sibling 1's outcome*/
gen y_2=. /*sibling 2's outcome*/
gen d=. /*gain-score (y_2-y_1)*/

**simulation

tempname m2c
postfile `m2c' sc_mc6 sc_ub_mc6 sc_lb_mc6 using m2c, replace
quietly {

forvalues i=1/$nmc {

   /*note: for y_1 and y_2 equations, "rnormal(0,1)" is residual*/

   replace u=rnormal(0,1)
   replace t_1=(chi*u)>0.5
   replace t_2=(phi*t_1)+(gamma*u)>0.2
   replace y_1=(delta*t_1)+(kappa*t_2)+(psi*u)+rnormal(0,1)
   replace y_2=(delta*t_2)+(theta*t_1)+(psi*u)+rnormal(0,1)
   replace d=y_2-y_1

   /*gain-score regression*/
   reg d t_1 t_2

   /*lincom to compute spillover coefficient*/
   lincom t_1+t_2

```

```
    /*store lincom output*/
    scalar sc_mc6=r(estimate) /*spillover coefficient*/
    scalar sc_ub_mc6=(r(estimate)+(1.96*r(se))) /*95% CI upper bound*/
    scalar sc_lb_mc6=(r(estimate)-(1.96*r(se))) /*95% CI lower bound*/

    post `m2c' (sc_mc6) (sc_ub_mc6) (sc_lb_mc6)
 }
 }
postclose `m2c'

use m2c, clear
summarize

**store mean of results, save to append with other simulation results
egen sc=mean(sc_mc6)
egen ub=mean(sc_ub_mc6)
egen lb=mean(sc_lb_mc6)
gen id=6 /*sixth simulation*/

keep sc ub lb id
duplicates drop
duplicates report

**save data
save "sim6.dta", replace


/*********************/
*Simulate Model 3A
/*********************/

**clear previous data
clear all
set more off

**monte carlo simulation settings
global nobs=5000 /*observations per run*/
global nmc=1000 /*number of runs*/
set seed 278 /*seed for random number generator -- can change*/
set obs $nobs /*sets observations within each run*/

**set parameters (causal effects in models)
scalar delta=1 /*t_1 on y_1, t_2 on y_2*/

scalar psi=1 /*u on y_1, u on y_2*/
scalar chi=2 /*u on t_1*/
scalar gamma=3 /*u on t_2*/

scalar theta=0.5 /*t_1 on y_2*/
scalar kappa=0.3 /*t_2 on y_1*/
```

```
scalar tau=0.3 /*t_2 on t_1*/
scalar phi=0.3 /*t_1 on t_2*/

scalar omega=0.3 /*y_1 on t_2*/
scalar eta=0.3 /*y_1 on y_2*/
scalar lambda=0.3 /*y_2 on y_1*/

**generate variables

gen u=. /*unobserved fixed effect*/
gen t_1=. /*sibling 1's treatment*/
gen t_2=. /*sibling 2's treatment*/
gen y_1=. /*sibling 1's outcome*/
gen y_2=. /*sibling 2's outcome*/
gen d=. /*gain-score (y_2-y_1)*/

**simulation

tempname m3a
postfile `m3a' sc_mc7 sc_ub_mc7 sc_lb_mc7 using m3a, replace
quietly {

forvalues i=1/$nmc {

   /*note: for y_1 and y_2 equations, "rnormal(0,1)" is residual*/

   replace u=rnormal(0,1)
   replace t_1=(chi*u)>0.5
   replace y_1=(delta*t_1)+(psi*u)+rnormal(0,1)
   replace t_2=(omega*y_1)+(gamma*u)>0.2
   replace y_2=(delta*t_2)+(theta*t_1)+(psi*u)+rnormal(0,1)
   replace d=y_2-y_1

   /*gain-score regression*/
   reg d t_1 t_2

   /*lincom to compute spillover coefficient*/
   lincom t_1+t_2

   /*store lincom output*/
   scalar sc_mc7=r(estimate) /*spillover coefficient*/
   scalar sc_ub_mc7=(r(estimate)+(1.96*r(se))) /*95% CI upper bound*/
   scalar sc_lb_mc7=(r(estimate)-(1.96*r(se))) /*95% CI lower bound*/

   post `m3a' (sc_mc7) (sc_ub_mc7) (sc_lb_mc7)
}
}
postclose `m3a'

use m3a, clear
```

```
summarize

**store mean of results, save to append with other simulation results
egen sc=mean(sc_mc7)
egen ub=mean(sc_ub_mc7)
egen lb=mean(sc_lb_mc7)
gen id=7 /*seventh simulation*/

keep sc ub lb id
duplicates drop
duplicates report

**save data
save "sim7.dta", replace


/*********************/
*Simulate Model 3B
/*********************/

**clear previous data
clear all
set more off

**monte carlo simulation settings
global nobs=5000 /*observations per run*/
global nmc=1000 /*number of runs*/
set seed 53893 /*seed for random number generator -- can change*/
set obs $nobs /*sets observations within each run*/

**set parameters (causal effects in models)
scalar delta=1 /*t_1 on y_1, t_2 on y_2*/

scalar psi=1 /*u on y_1, u on y_2*/
scalar chi=2 /*u on t_1*/
scalar gamma=3 /*u on t_2*/

scalar theta=0.5 /*t_1 on y_2*/
scalar kappa=0.3 /*t_2 on y_1*/
scalar tau=0.3 /*t_2 on t_1*/
scalar phi=0.3 /*t_1 on t_2*/

scalar omega=0.3 /*y_1 on t_2*/
scalar eta=0.3 /*y_1 on y_2*/
scalar lambda=0.3 /*y_2 on y_1*/

**generate variables

gen u=. /*unobserved fixed effect*/
gen t_1=. /*sibling 1's treatment*/
```

```
gen t_2=. /*sibling 2's treatment*/
gen y_1=. /*sibling 1's outcome*/
gen y_2=. /*sibling 2's outcome*/
gen d=. /*gain-score (y_2-y_1)*/

**simulation

tempname m3b
postfile `m3b' sc_mc8 sc_ub_mc8 sc_lb_mc8 using m3b, replace
quietly {

forvalues i=1/$nmc {

   /*note: for y_1 and y_2 equations, "rnormal(0,1)" is residual*/

   replace u=rnormal(0,1)
   replace t_1=(chi*u)>0.5
   replace t_2=(gamma*u)>0.2
   replace y_1=(delta*t_1)+(psi*u)+rnormal(0,1)
   replace y_2=(delta*t_2)+(theta*t_1)+(eta*y_1)+(psi*u)+rnormal(0,1)
   replace d=y_2-y_1

   /*gain-score regression*/
   reg d t_1 t_2

   /*lincom to compute spillover coefficient*/
   lincom t_1+t_2

   /*store lincom output*/
   scalar sc_mc8=r(estimate) /*spillover coefficient*/
   scalar sc_ub_mc8=(r(estimate)+(1.96*r(se))) /*95% CI upper bound*/
   scalar sc_lb_mc8=(r(estimate)-(1.96*r(se))) /*95% CI lower bound*/

   post `m3b' (sc_mc8) (sc_ub_mc8) (sc_lb_mc8)
}
}
postclose `m3b'

use m3b, clear
summarize

**store mean of results, save to append with other simulation results
egen sc=mean(sc_mc8)
egen ub=mean(sc_ub_mc8)
egen lb=mean(sc_lb_mc8)
gen id=8 /*eighth simulation*/

keep sc ub lb id
duplicates drop
duplicates report
```

```
**save data
save "sim8.dta", replace


/*********************/
*Simulate Model 3C
/*********************/

**clear previous data
clear all
set more off

**monte carlo simulation settings
global nobs=5000 /*observations per run*/
global nmc=1000 /*number of runs*/
set seed 9921 /*seed for random number generator -- can change*/
set obs $nobs /*sets observations within each run*/

**set parameters (causal effects in models)
scalar delta=1 /*t_1 on y_1, t_2 on y_2*/

scalar psi=1 /*u on y_1, u on y_2*/
scalar chi=2 /*u on t_1*/
scalar gamma=3 /*u on t_2*/

scalar theta=0.5 /*t_1 on y_2*/
scalar kappa=0.3 /*t_2 on y_1*/
scalar tau=0.3 /*t_2 on t_1*/
scalar phi=0.3 /*t_1 on t_2*/

scalar omega=0.3 /*y_1 on t_2*/
scalar eta=0.3 /*y_1 on y_2*/
scalar lambda=0.3 /*y_2 on y_1*/

**generate variables

gen u=. /*unobserved fixed effect*/
gen t_1=. /*sibling 1's treatment*/
gen t_2=. /*sibling 2's treatment*/
gen y_1=. /*sibling 1's outcome*/
gen y_2=. /*sibling 2's outcome*/
gen d=. /*gain-score (y_2-y_1)*/

**simulation

tempname m3c
postfile `m3c' sc_mc9 sc_ub_mc9 sc_lb_mc9 using m3c, replace
quietly {

```

```
forvalues i=1/$nmc {

   /*note: for y_1 and y_2 equations, "rnormal(0,1)" is residual*/

   replace u=rnormal(0,1)
   replace t_1=(chi*u)>0.5
   replace t_2=(gamma*u)>0.2
   replace y_2=(delta*t_2)+(theta*t_1)+(psi*u)+rnormal(0,1)
   replace y_1=(delta*t_1)+(lambda*y_2)+(psi*u)+rnormal(0,1)
   replace d=y_2-y_1

   /*gain-score regression*/
   reg d t_1 t_2

   /*lincom to compute spillover coefficient*/
   lincom t_1+t_2

   /*store lincom output*/
   scalar sc_mc9=r(estimate) /*spillover coefficient*/
   scalar sc_ub_mc9=(r(estimate)+(1.96*r(se))) /*95% CI upper bound*/
   scalar sc_lb_mc9=(r(estimate)-(1.96*r(se))) /*95% CI lower bound*/

   post `m3c' (sc_mc9) (sc_ub_mc9) (sc_lb_mc9)
}
}
postclose `m3c'

use m3c, clear
summarize

**store mean of results, save to append with other simulation results
egen sc=mean(sc_mc9)
egen ub=mean(sc_ub_mc9)
egen lb=mean(sc_lb_mc9)
gen id=9 /*ninth simulation*/

keep sc ub lb id
duplicates drop
duplicates report

**save data
save "sim9.dta", replace


/*********************/
*Graph Results
/*********************/

clear
set more off
```

```
**open first simulation dataset
use "sim1.dta"

**append datasets (sim2-9)
append using "sim2.dta"
append using "sim3.dta"
append using "sim4.dta"
append using "sim5.dta"
append using "sim6.dta"
append using "sim7.dta"
append using "sim8.dta"
append using "sim9.dta"

**label id (assign to models)
label define model_lbl 1 "Figure 2.1A" 2 "1B" 3 "1C" ///
   4 "Figure 2.2A" 5 "2B" 6 "2C" ///
   7 "Figure 2.3A" 8 "3B" 9 "3C"

label values id model_lbl

**display results

/*
NOTE:
id = sibling spillover model
sc = spillover coefficient
lb = 95% CI lower bound for spillover coefficient
ub = 95% CI upper bound for spillover coefficient
*/

list id sc lb ub

**graph results

/*
NOTE: When importing Stata code into LaTeX, the compiler will not accept the en-dash in "{&theta} [MINUS] {&kappa}." Replace the hyphen with an en-dash to recreate the original figure.
*/

twoway(scatter id sc, mcolor(black) ysc(reverse) ///
      graphregion(color(white)) ///
      xtitle(Mean Spillover Coefficient (95% CI)) ///
      ytitle(Sibling Spillover Model) ///
      text(8.5 0.5 "({&theta} = 0.5)", size(small) ///
         box margin(l+0.5 r+0.5 t+0.5 b+1)) ///
      text(8.5 0.2 "({&theta} - {&kappa} = 0.2)", size(small) ///
         box margin(l+0.5 r+0.5 t+0.5 b+1)) ///
      xline(0.2, lcolor(black) lpattern(shortdash_dot)) ///
```

```
      xline(0.5, lcolor(black) lpattern(dash)) ///
      legend(off) xtick(#21)) ///
   (rcap ub lb id, horizontal lcolor(black)), ///
   ylabel(1(1)9, angle(horizontal) labsize(small) ///
      valuelabel notick) ///
   xlabel(-0.5(0.5)1.5, labsize(small)) saving(sim_results, replace)


/*

NOTE: The code below erases results (datasets and the figure). Highlight the
code below and press the "Execute (do)" button to delete files.

**erase datasets with simulation results
erase m1a.dta
erase m1b.dta
erase m1c.dta
erase m2a.dta
erase m2b.dta
erase m2c.dta
erase m3a.dta
erase m3b.dta
erase m3c.dta

erase sim1.dta
erase sim2.dta
erase sim3.dta
erase sim4.dta
erase sim5.dta
erase sim6.dta
erase sim7.dta
erase sim8.dta
erase sim9.dta

**erase figure with simulation results
erase sim_results.gph
```

References

[1] Pearl J. Linear models: a useful ”microscope” for causal analysis. *J Causal Inference*. 2013;1:155-170.

[2] Invernizzi M, Juel C, Swank L, Meier J. PALS-K Technical Reference. Charlottesville, VA: University of Virginia Curry School of Education, 2015.

[3] Larson A, Berger LM, Mallinson DC, Grodsky E, Ehrenthal DB. Variable uptake of Medicaid-covered Prenatal Care Coordination: the relevance of treatment level and service context. *J Community Health*. 2019;44:32-43.

[4] Mallinson DC, Grodsky E, Ehrenthal DB. Gestational age, kindergarten-level literacy, and effect modification by maternal socio-economic and demographic factors. *Paediatr Perinat Epidemiol*. 2019;33:467-479.

[5] Wisconsin Department of Public Instruction. Admissions and early entrance to four- and five-year-old kindergarten. https://dpi.wi.gov/early-childhood/kind/admission (1 September 2020, date last accessed).

[6] Ford K, Invernizzi M. PALS Español K Technical Reference. Charlottesville, VA: University of Virginia Curry School of Education; 2014.

[7] StataCorp. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC; 2019.