Online methods for

“**A simple approximation to bias in gene-environment interaction estimates when a case might not be the case**”

Iryna Lobach, Inyoung Kim, Alexander Alekseyenko, Siarhei Lobach, Li Zhang

July 11, 2019

**APPENDIX**

**DERIVATION OF THE APPROXIMATION**

The observed data are collected using retrospective case-control sampling design, i.e. the genetic and environmental variables are sampled conditionally on the clinical diagnosis. The likelihood function of the observed data is based on the probability and hence we define . (4)

The usual logistic regression model with the clinically diagnosed disease status as the outcome variable is

, (5)

We are interested to find an analytic solution that relates parameters from the misspecified model (5) to the parameters and from the *true* model (4).

The next steps are motivated by the developments in Kullback (1959), Neuhaus (1999). Kullback (1959) showed that parameters estimated in the misspesified model (4) converge to values that minimize the Kullback-Leibler divergence between the *true* and false models with expectations taken with respect to the *true* model, i.e.

. (5)

We first examine the true model (1) and misspecified model (3).

We first derive an approximation to the relationship between the parameters of the misspecified model (3) to the parameters of the true model (1).

Derivative of the Kullback-Leibler divergence (5) with respect to parameters of the misspecified model are

Define , then taking derivatives of Kullback-Leibler divergence (5) with respect to we arrive to the following system of equations

; (A1)

;

;

Values of such that

(A2)

for all solve the system of equations (A2). Hence values of for which and for any solve the system of equations (A2).

We first consider a setting with . By definition, . Because values of for which solve the system of equations (7),

= . (A3)

In the next equations we apply the same arguments for the other parameters.

; (A4)

; (A5)

. (A6)

While the equations (A2)-(A6) describe the relationship between the parameters and , it is not clear what the relationship is. Taylor series expansion around 0 provides convenient approximation in the form

(A7)

; (A8)

(A9)

. (A10)

It can be easily seen that when a variable is included in the model,

; (A11)

(A12)

; (A13)

(A14)

. (A15)

With Taylor series expansion the system of equations becomes (A11)-(A15)

; (A11.1)

; (A12.1)

; (A13.1)

; (A14.1)

(A15.1)

We next suppose that the true model is (2), while the parameters are estimated based on a misspecified model (3).

By definition, . Because values of parameters for which minimize the Kullback-Leibler divergence (5), . Algebraic derivations show that then and Taylor series expansion around zero arrives at an approximation Similar derivations apply to the other parameters:

(A16)

(A17)

(A18)

When an environmental variable that does not interact with the genotype is in the model, then the formulas are as follows:

When an environmental variable that does not interact with the genotype is in the model, then the formulas are as follows:

; (A19)

(A20)

(A21)

(A22)

(A23)

With Taylor series expansion the system of equations (A19)-(A23) becomes

; (A19.1)

(A20.1)

(A21.1)

(A22.1)

(A23.1)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| True model (1) | | Misspecified Model (3) | | | |
| Parameter | True Value | Parameter | Empirical average  estimate | Empirical SD | Approximation (A11)-(A15) |
|  | 0.41 |  | 0.28 | 0.10 | 0.26 |
|  | 2.08 |  | 1.80 | 0.09 | 1.30 |
|  | 0 |  | 0.13 | 0.36 | 0.11 |
|  | 0 |  | -0.00 | 0.05 | 0 |
|  | -0.08 |  | -0.06 | 0.05 | -0.06 |

**Supplementary Table 1.** Estimates of coefficients obtained empirically as an average across 500 simulated datasets with 3,000 cases and 3,000 controls. The genetic variable is Bernoulli with frequency 0.1. The environmental variables are Bernoulli with frequencies 14%, 50%, and 52%, respectively. The *true* diagnosis is simulated based on the model (1) with coefficients and The clinical-pathological diagnoses relationship is defined as and . Coefficients s are estimated based on misspecified model (3).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| True model (2) | | Misspecified Model (3) | | | |
| Parameter | True Value | Parameter | Empirical average | Empirical SD | Approximation (A16)-(A22) |
|  | -0.69 |  | -0.38 | 0.10 | -0.39 |
|  | 1.3 |  | 1.08 | 0.08 | 1.08 |
|  | 0.10 |  | -0.10 | 0.25 | -0.05 |
|  | 0.10 |  | 0.10 | 0.05 | 0.10 |
|  | -0.08 |  | -0.09 | 0.05 | -0.08 |

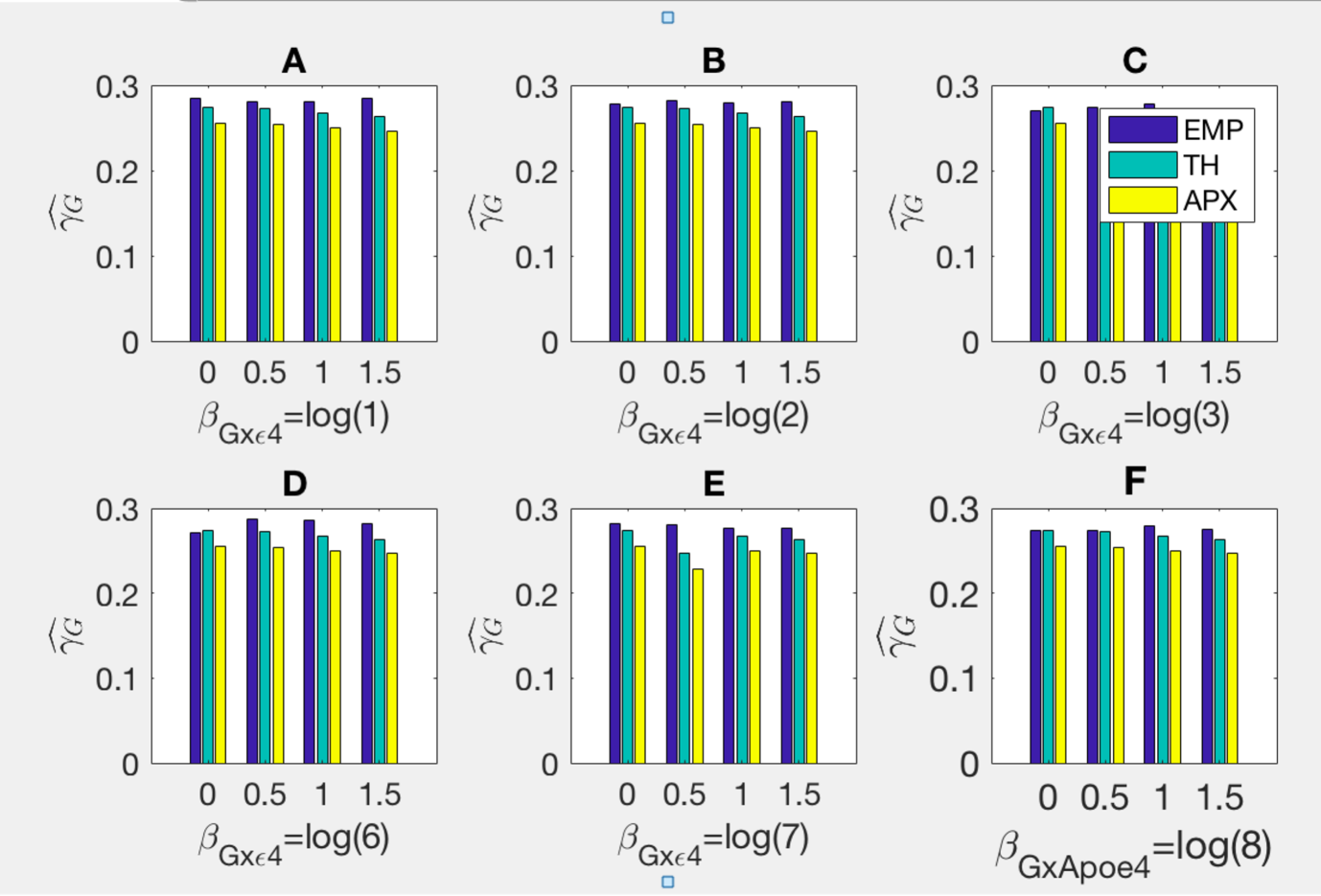
**Supplementary Table 2.** Estimates of coefficients obtained empirically as an average across 500 simulated datasets with 3,000 cases and 3,000 controls. Risk of the disease of interest is defined in a set of parameters ; while the risk of the nuisance disease is parametrized by Frequency of ApoE 4 allele in the population is 14%. Variables and are Bernoulli with frequencies 0.50 and 0.52, respectively. Frequencies of the disease of interest and the nuisance disease are Frequency of the nuisance disease within the clinical diagnosis varies by ApoE4 status )=0.36 and )=0.06.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| True model (2) | | Misspecified Model (3) | | | |
| Parameter | True Value | Parameter | Empirical average | Empirical SD | Approximation (A16)-(A22) |
|  | -0.69 |  | -0.33 | 0.14 | -0.39 |
|  | 1.3 |  | 1.12 | 0.12 | 1.08 |
|  | 0.10 |  | -0.10 | 0.30 | -0.05 |
|  | 0.10 |  | 0.10 | 0.11 | 0.10 |
|  | -0.08 |  | -0.03 | 0.12 | -0.08 |

**Supplementary Table 3.** Estimates of coefficients obtained empirically as an average across 500 simulated datasets with 2,000 cases and 1,000 controls. Risk of the disease of interest is defined in a set of parameters ; while the risk of the nuisance disease is parametrized by Frequency of ApoE 4 allele in the population is 14%. Variables and are Bernoulli with frequencies 0.50 and 0.52, respectively. Frequencies of the disease of interest and the nuisance disease are Frequency of the nuisance disease within the clinical diagnosis varies by ApoE4 status )=0.36 and )=0.06.

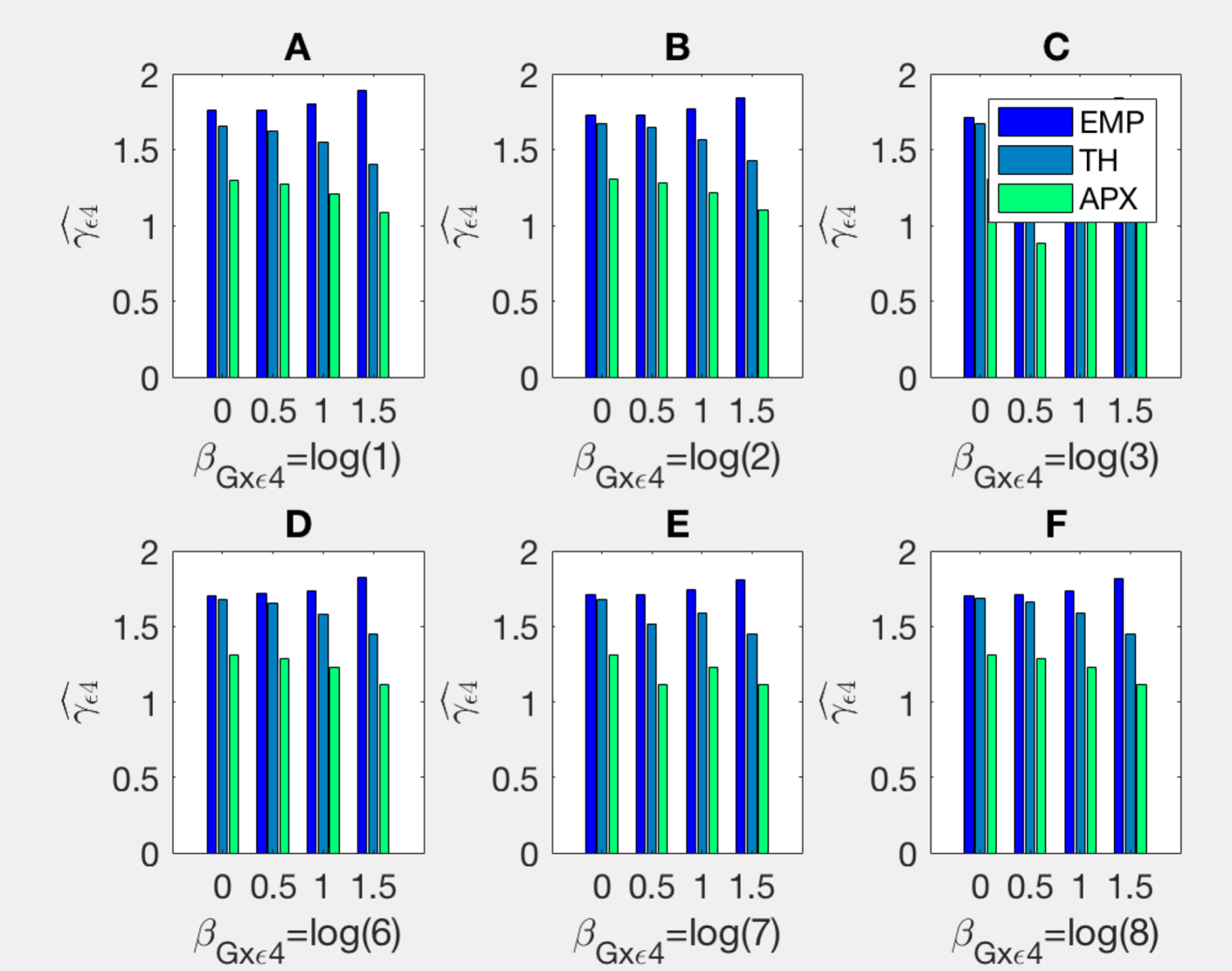
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| True model (2) | | Misspecified Model (3) | | | | | | | |
| Parameter | Approximation (A16)-(A22) | Rates of nuisance disease state are overestimated for both carriers and non-carriers | | Rates of nuisance disease state are underestimated for both carriers and non-carriers | | Rates of nuisance disease state are overestimated for carriers and underestimated fornon-carriers | |
| Parameter | True Value | Empirical average | Empirical SD | Empirical average | Empirical SD | Empirical average | Empirical SD |
|  | -0.69 |  | -0.39 | -0.34 | 0.17 | -0.37 | 0.16 | -0.39 | 0.20 |
|  | 1.3 |  | 1.08 | 1.14 | 0.14 | 1.12 | 0.15 | 1.13 | 0.19 |
|  | 0.10 |  | -0.05 | -0.08 | 0.33 | -0.11 | 0.33 | -0.11 | 0.39 |
|  | 0.10 |  | 0.10 | 0.10 | 0.07 | 0.11 | 0.09 | 0.10 | 0.10 |
|  | -0.08 |  | -0.08 | -0.05 | 0.10 | -0.09 | 0.12 | -0.12 | 0.12 |

**Supplementary Table 3.** Estimates of coefficients obtained empirically as an average across 500 simulated datasets with 2,000 cases and 1,000 controls. Risk of the disease of interest is defined in a set of parameters ; while the risk of the nuisance disease is parametrized by Frequency of ApoE 4 allele in the population is 14%. Variables and are Bernoulli with frequencies 0.50 and 0.52, respectively. Frequencies of the disease of interest and the nuisance disease are Frequency of the nuisance disease within the clinical diagnosis varies by ApoE4 status )=0.36 and )=0.06.

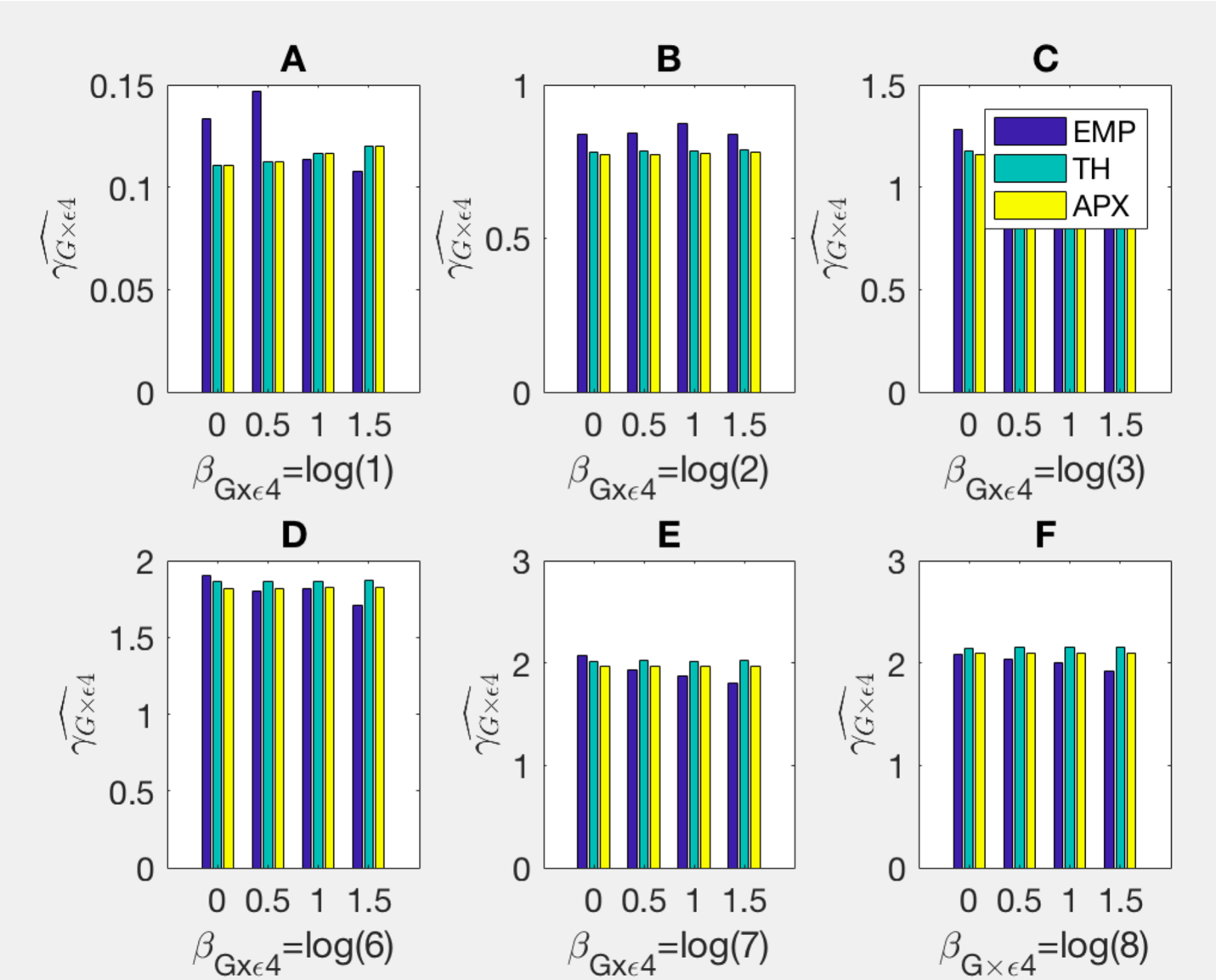


**Supplementary Figure 1A**: Estimates of obtained empirically as an average across 500 simulated datasets (EMP) with 3,000 cases and 3,000 controls; using the theoretical formulae (A11)-(A15) (TH); and using the approximation (A11.1)-(A15.1) (APX). Coefficients s are estimated based on misspecified model (3).

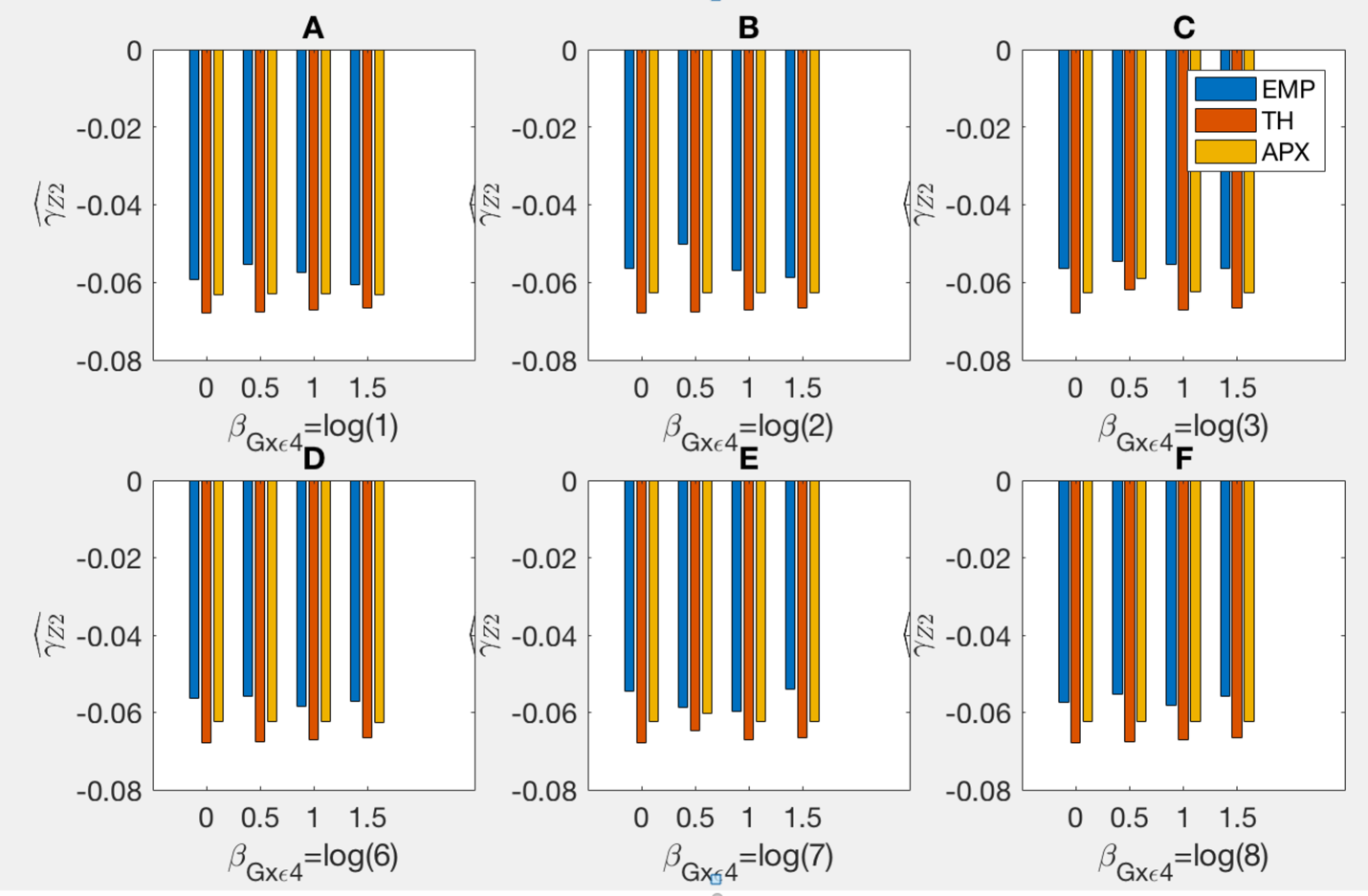
The genetic variable is Bernoulli with frequency 0.1. The environmental variables are Bernoulli with frequencies 14%, 50%, and 52%, respectively. The *true* diagnosis is simulated based on the model (1) with coefficients and The clinical-pathological diagnoses relationship is defined as and .



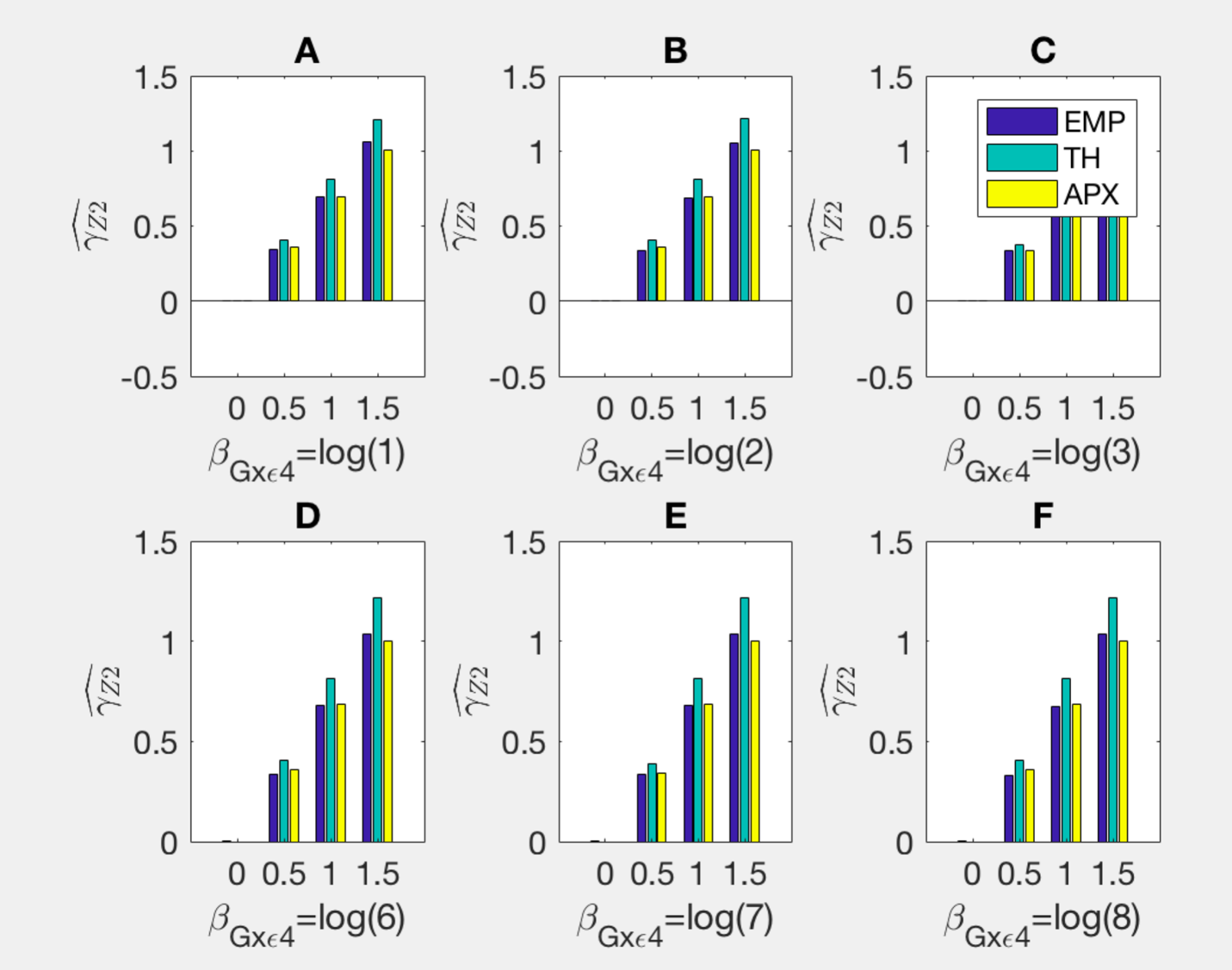
**Supplementary Figure 1B**: Estimates of obtained empirically as an average across 500 simulated datasets (EMP); with 3,000 cases and 3,000 controls; using the theoretical formulae (A11)-(A15) as the difference in logs (TH); and using the approximation (A11)-(A15) (APX). Coefficients s are estimated based on misspecified model (3). The genetic variable is Bernoulli with frequency 0.1. The environmental variables are Bernoulli with frequencies 14%, 50%, and 52%, respectively. The *true* diagnosis is simulated based on the model (1) with coefficients and The clinical-pathological diagnoses relationship is defined as and .



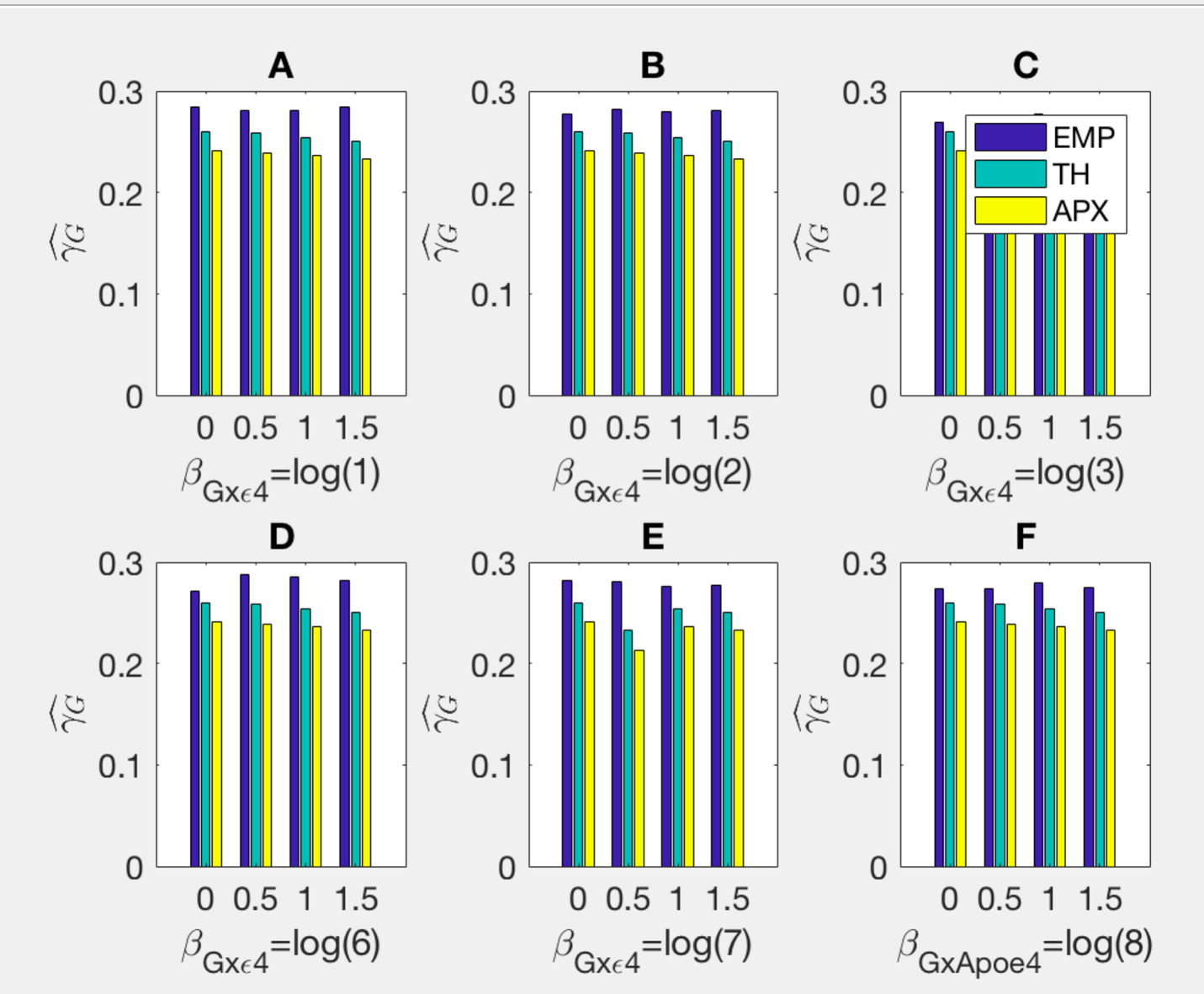
**Supplementary Figure 1C**: Estimates of obtained empirically as an average across 500 simulated datasets (EMP); with 3,000 cases and 3,000 controls; using the theoretical formulae (A11)-(A15) as the difference in logs (TH); and using the approximation (A11)-(A15) (APX). Coefficients s are estimated based on misspecified model (3). The genetic variable is Bernoulli with frequency 0.1. The environmental variables are Bernoulli with frequencies 14%, 50%, and 52%, respectively. The *true* diagnosis is simulated based on the model (1) with coefficients and The clinical-pathological diagnoses relationship is defined as and .



**Supplementary Figure 1D**: Estimates of obtained empirically as an average across 500 simulated datasets (EMP); with 3,000 cases and 3,000 controls; using the theoretical formulae (A11)-(A15) as the difference in logs (TH); and using the approximation (A11)-(A15) (APX). Coefficients s are estimated based on misspecified model (3). The genetic variable is Bernoulli with frequency 0.1. The environmental variables are Bernoulli with frequencies 14%, 50%, and 52%, respectively. The *true* diagnosis is simulated based on the model (1) with coefficients and The clinical-pathological diagnoses relationship is defined as and .

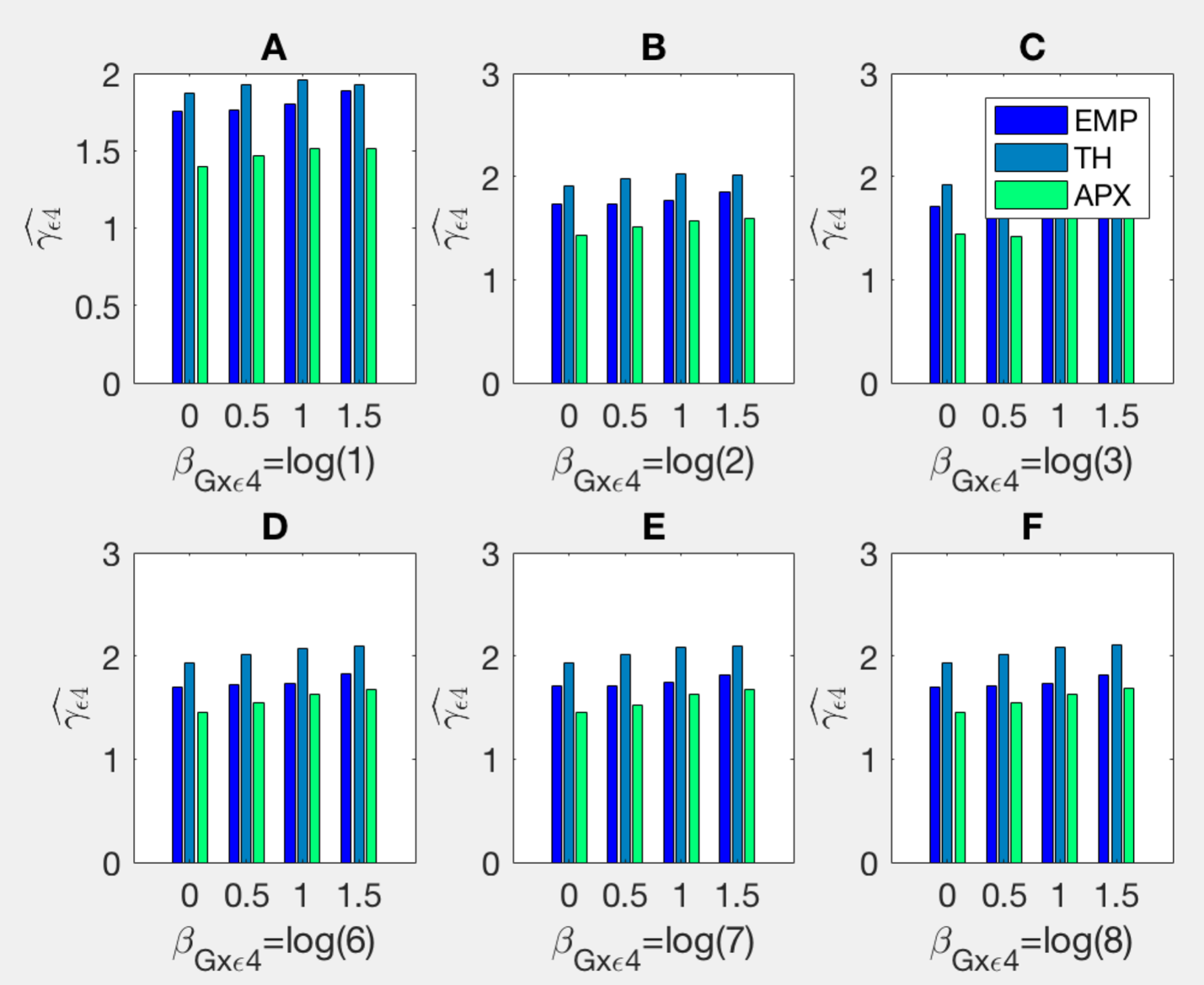
****

**Supplementary Figure 1E**: Estimates of obtained empirically as an average across 500 simulated datasets (EMP) with 3,000 cases and 3,000 controls; using the theoretical formulae (A11)-(A15) as the difference in logs (TH); and using the approximation (A11)-(A15) (APX). Coefficients s are estimated based on misspecified model (3). The genetic variable is Bernoulli with frequency 0.1. The environmental variables are Bernoulli with frequencies 14%, 50%, and 52%, respectively. The *true* diagnosis is simulated based on the model (1) with coefficients and The clinical-pathological diagnoses relationship is defined as and .

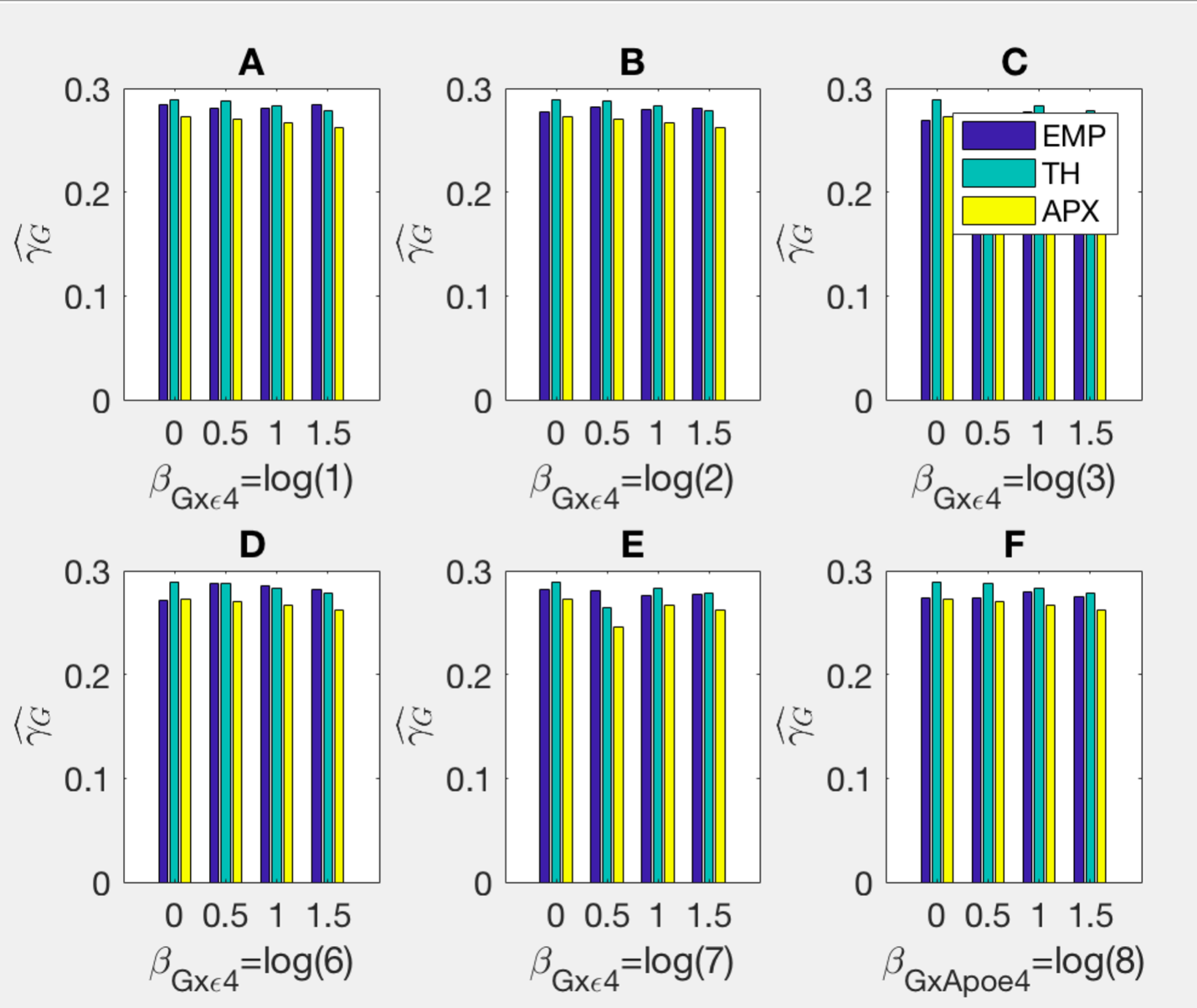


**Web Supplementary Figure 2A**: Estimates of obtained empirically as an average across 500 simulated datasets (EMP) with 3,000 cases and 3,000 controls; using the theoretical formulae (A11)-(A15) (TH); and using the approximation (A11.1)-(A15.1) (APX). Coefficients s are estimated based on misspecified model (3).

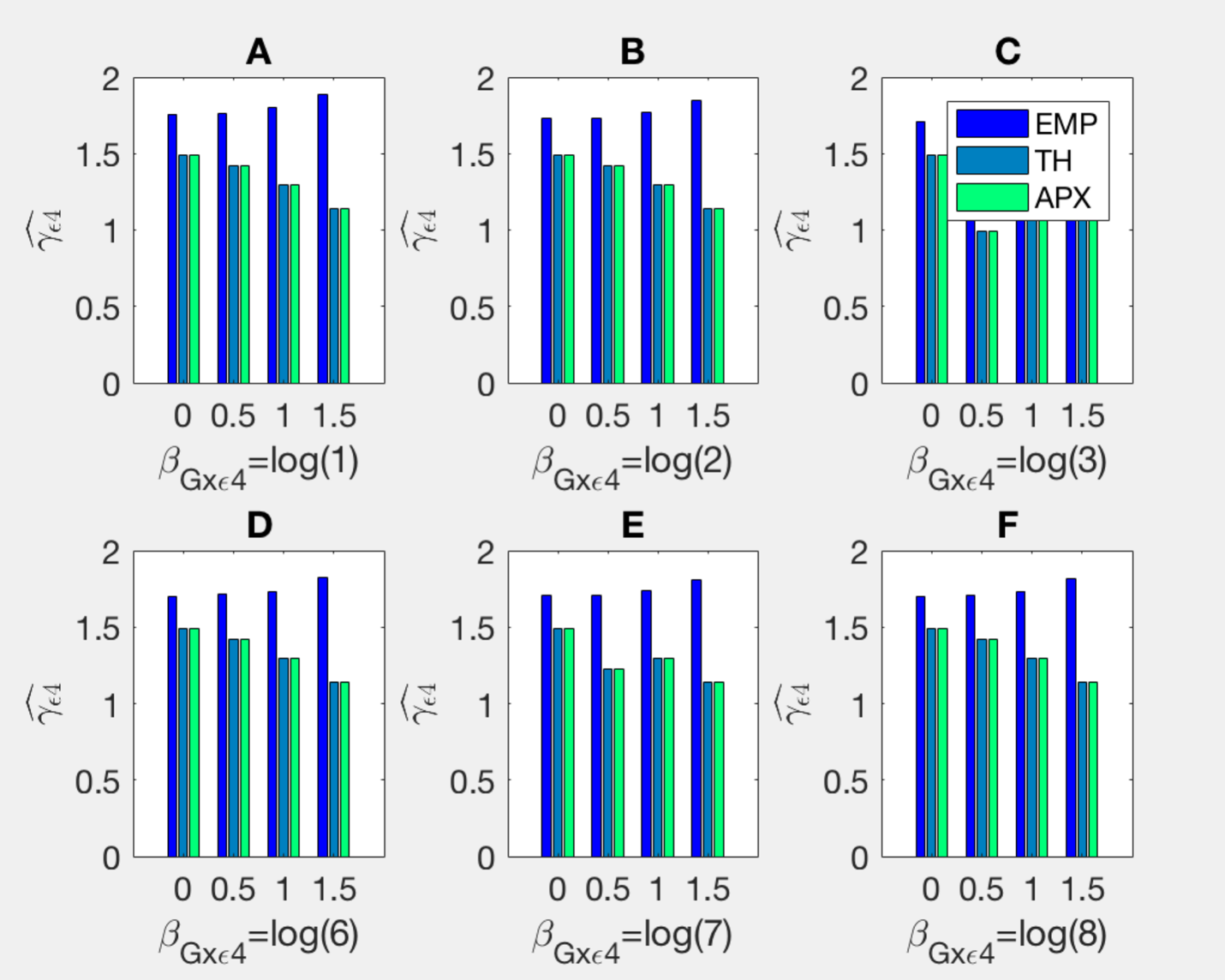
The genetic variable is Bernoulli with frequency 0.1. The environmental variables are Bernoulli with frequencies 14%, 50%, and 52%, respectively. The *true* diagnosis is simulated based on the model (1) with coefficients and The clinical-pathological diagnoses relationship is defined as and , but in the theoretical derivations is underestimated to be and .



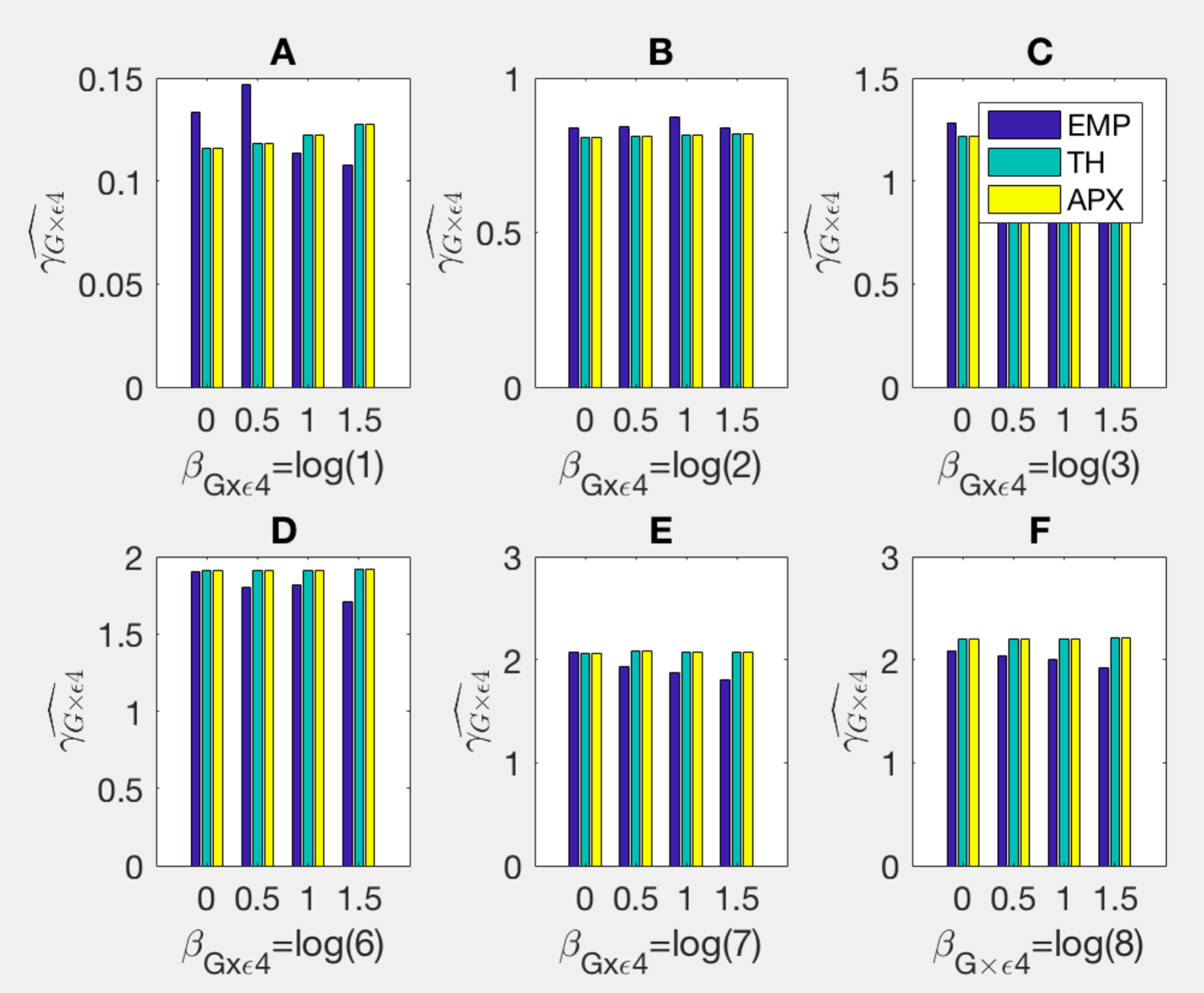
**Web Supplementary Figure 2B**: Estimates of obtained empirically as an average across 500 simulated datasets (EMP); with 3,000 cases and 3,000 controls; using the theoretical formulae (A11)-(A15) as the difference in logs (TH); and using the approximation (A11)-(A15) (APX). Coefficients s are estimated based on misspecified model (3). The genetic variable is Bernoulli with frequency 0.1. The environmental variables are Bernoulli with frequencies 14%, 50%, and 52%, respectively. The *true* diagnosis is simulated based on the model (1) with coefficients and The clinical-pathological diagnoses relationship is defined as and , but in the theoretical derivations is underestimated to be and .



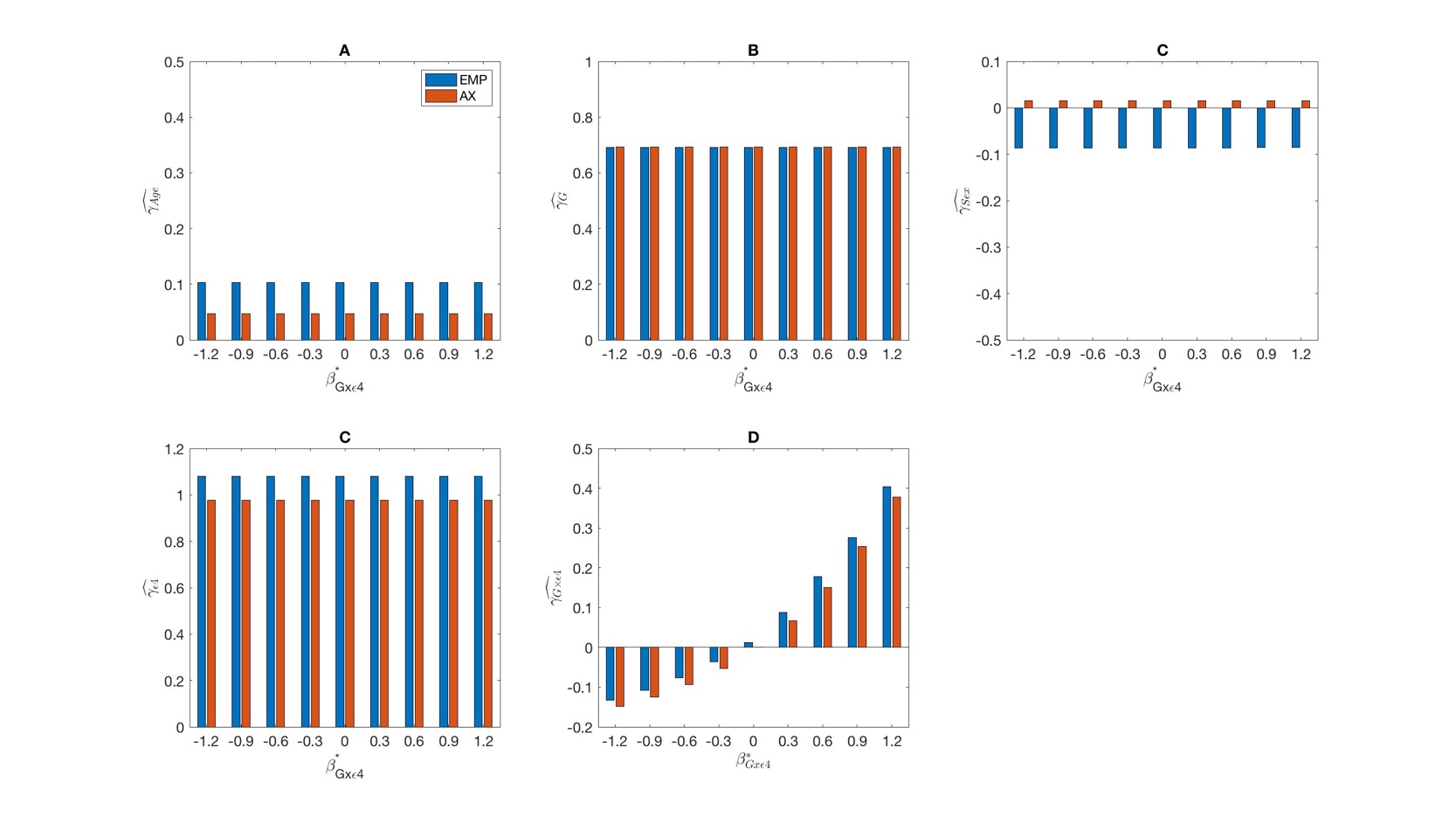
**Web Supplementary Figure 2C**: Estimates of obtained empirically as an average across 500 simulated datasets (EMP); with 3,000 cases and 3,000 controls; using the theoretical formulae (A11)-(A15) as the difference in logs (TH); and using the approximation (A11)-(A15) (APX). Coefficients s are estimated based on misspecified model (3). The genetic variable is Bernoulli with frequency 0.1. The environmental variables are Bernoulli with frequencies 14%, 50%, and 52%, respectively. The *true* diagnosis is simulated based on the model (1) with coefficients and The clinical-pathological diagnoses relationship is defined as and , but in the theoretical derivations is underestimated to be and .



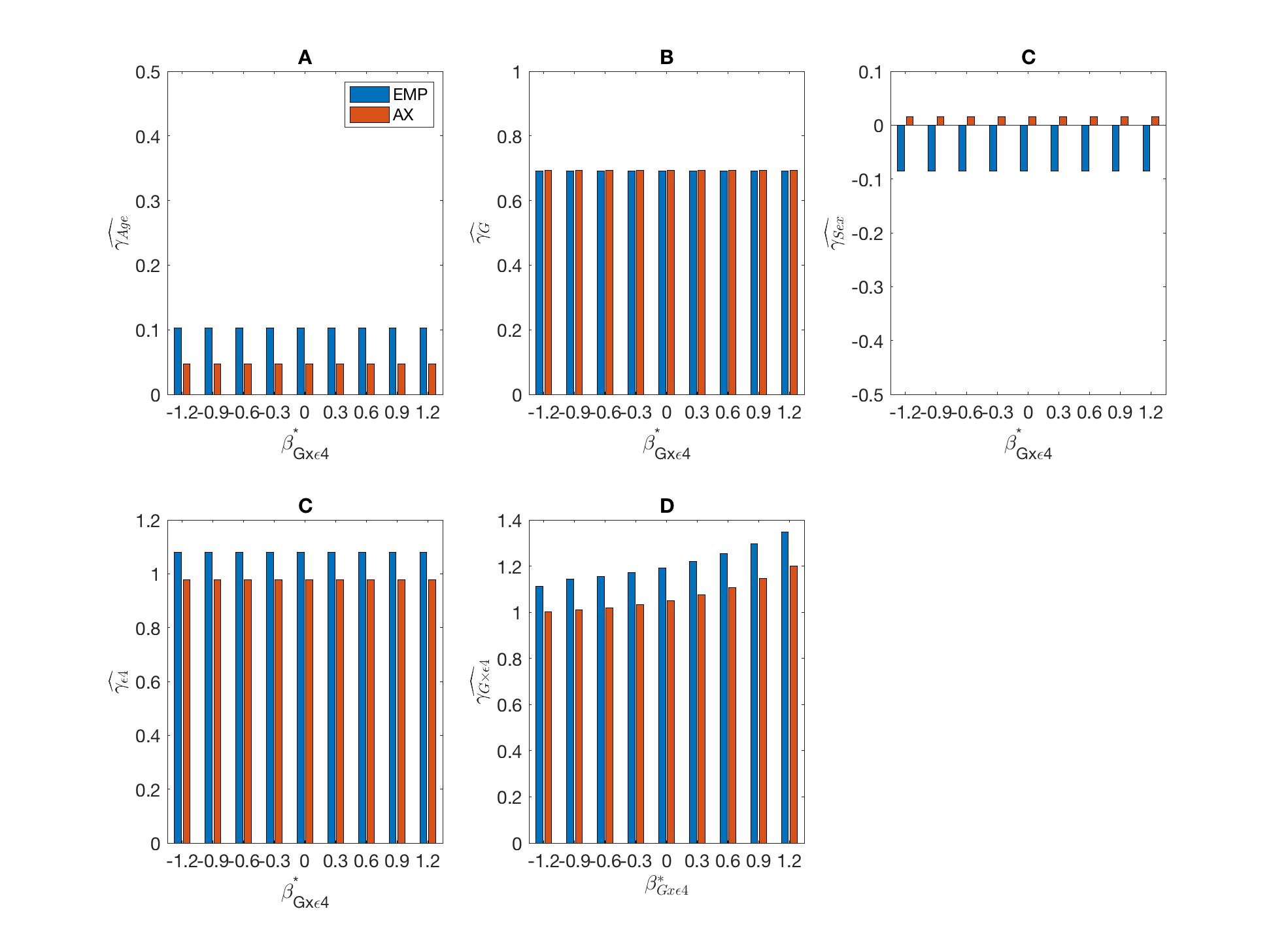
**Web Supplementary Figure 3A**: Estimates of obtained empirically as an average across 500 simulated datasets (EMP); with 3,000 cases and 3,000 controls; using the theoretical formulae (A11)-(A15) as the difference in logs (TH); and using the approximation (A11)-(A15) (APX). Coefficients s are estimated based on misspecified model (3). The genetic variable is Bernoulli with frequency 0.1. The environmental variables are Bernoulli with frequencies 14%, 50%, and 52%, respectively. The *true* diagnosis is simulated based on the model (1) with coefficients and The clinical-pathological diagnoses relationship is defined as and , but in the theoretical derivations is overestimated to be and .



**Web Supplementary Figure 3B**: Estimates of obtained empirically as an average across 500 simulated datasets (EMP); with 3,000 cases and 3,000 controls; using the theoretical formulae (A11)-(A15) as the difference in logs (TH); and using the approximation (A11)-(A15) (APX). Coefficients s are estimated based on misspecified model (3). The genetic variable is Bernoulli with frequency 0.1. The environmental variables are Bernoulli with frequencies 14%, 50%, and 52%, respectively. The *true* diagnosis is simulated based on the model (1) with coefficients and The clinical-pathological diagnoses relationship is defined as and , but in the theoretical derivations is underestimated to be and .



**Web Supplementary Figure 4A**: Empirical estimates (EMP) and approximation (AX) of parameter estimates when the data are simulated according to the risk model (2) while parameters are estimated based on the model (3). EMPs are the averages across 500 simulated datasets with 3,000 cases and 3,000 controls. Approximations are obtained using (A20)-(A23). We simulate the disease status *D*=1 vs. *D*=0 based on parameters; and we simulate *D*= vs. *D*=0 using with main effects of (Age) and (Sex) that are the same as for *D*=1 vs. *D*=0.



**Web Supplementary Figure 4B**: Empirical estimates (EMP) and approximation (AX) of parameter estimates when the data are simulated according to the risk model (2) while parameters are estimated based on the model (3). EMPs are the averages across 500 simulated datasets with 3,000 cases and 3,000 controls. Approximations are obtained using (A20)-(A23). We simulate the disease status *D*=1 vs. *D*=0 based on parameters and we simulate *D*= vs. *D*=0 using with main effects of (Age) and (Sex) that are the same as for *D*=1 vs. *D*=0.