Supplementary Material 3
Additional simulation experiment to investigate inflated type I errors

During the simulation study, the model-integrated method with covariance matrix and bootstrap for uncertainty showed inflated type I error for certain scenarios. Here we carried out further investigations regarding the chance of observing a high type I error using the current simulation scaffold.

We repeated the simulation study with 500 datasets 3 more times with different seeds used to generate the datasets. Considering the running time, we only analyzed the datasets using the two methods that had the shortest running times (the model-integrated method with covariance matrix assessing uncertainty and the standard NCA method). Further, we only investigated one of the study designs investigated in the main manuscript (24 individuals, 10 samples per individual, using the model with high variation). Figure S3-1 shows that for the additional 3 simulations (split_2, split_3, and split_4), the type I error of the covariance matrix method stayed within the expected interval (3.2-7.0%), indicating the type I error was not different from 5% for each repetition. However, if we combined the results from all the four simulation experiments, the type I error of the covariance matrix method was significantly different from 5% (Figure S3-2).

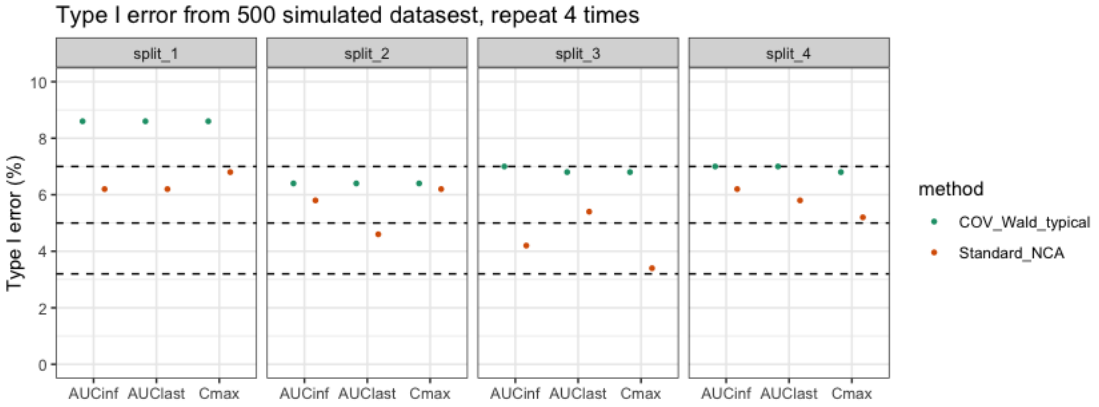

Figure S3-1: Type I errors when a simulation study with 500 datasets was performed for 4 times.

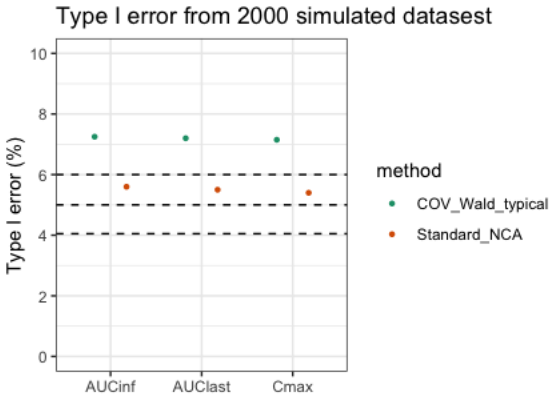

Figure S3-2: Type I error with a simulation study with 2000 datasets.

To further investigate the power of the simulation scaffold (i.e., the power of detecting a type I error that is different from 5% given the truth that such a difference exists), power curves were calculated based on a binomial test. It is seen from Figure S3-3 that a simulation experiment with 500 datasets has lower power than that with 1000 and 2000 datasets. If the true type I error is 7% (approximately the type I error indicated in Figure S3-2), the power for 500 datasets was lower than 50% to identify that value different from 5%.

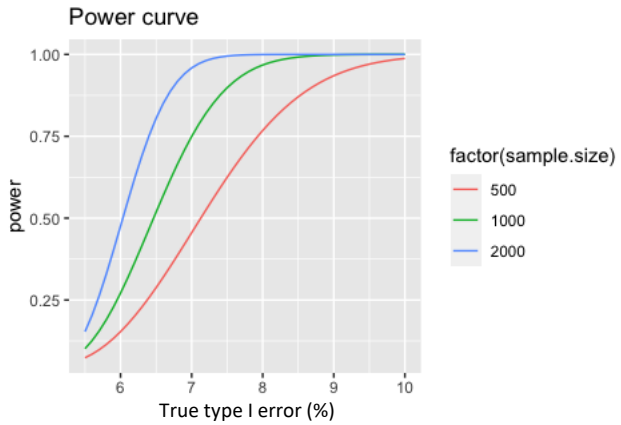

Figure S3-3: Power curves of simulation experiments with 500, 1000, and 2000 datasets.