Appendix: Software instructions to execute customized, study specific simulations

We provide an R package, simBA, to accompany this manuscript to allow users to execute simulations tailored to their study questions and needs.

Step 1: Install the latest GitLab release version of simBA using the devtools package via:

devtools::install_git("https://gitlab.partners.org/rjd48/unmeasured-confounding-simulations.git")

Step 2: Prepare an input table to base the data generation in the following format

Keep column headings unchanged. The table can be populated with user desired values for all columns. For binary variables being simulated, provide desired prevalence and for continuous variables, provide desired mean and standard deviation. Variables with empty cells next to them in the coeff_treatment_model and coeff_outcome_model columns will not be included in those models while generating the data. True confounders should have coefficients supplied for both these models in last two columns.

Variable	Description	Type	prevalence	mean	sd	coeff_treatment_model	coeff_outcome_model
c1	Gender	binary	0.21			-0.2783	0.46667
c2	Age	continuous		77	7.6	0.0321	0.07803
c3	GI complication history	binary	0.03			0.166	1.03895
c4	Concurrent GI protective agent use	binary	0.21			0.091	-0.68778
c5	warfarin use history	binary	0.08			0.5499	0.23121
c6	Number of hospitalizations	continuous		8	4.4	0.011	0.05539
c7	gastric ulcer history	binary	0.14			0.0488	
c8	OA diagnosis	binary	0.35			0.4507	
c9	RA diagnosis	binary	0.04			0.5488	
c10	GI protective agent use history	binary	0.28			0.1569	
c11	Any hospital admission in the CAP	binary	0.18			-0.0435	
c12	COPD	binary	0.15				0.24222
c13	Corticosteroid use history	binary	0.07				-0.13412
u1	unmeasured	binary	0.1			1	1

Step 3: Supply arguments to the simBA_de_novo function to run simulation based bias analysis for unmeasured confounding

Following is the function call, all the arguments are required and described in detail below

```
Scenario1 <- simBA_de_novo(iterations=500,

parameter_file_path = "C:\\Users\\rjd48\\SimBA input table.xlsx",

size=5000, treatment_prevalence=0.25, treatment_coeff= -0.25,

outcome_prevalence=0.05, dist= 'E',

unmeasured_conf="u1", n_proxies=2, proxy_type='binary', corr=0.5)
```

Argument	**Details and possible values**
iterations	Number of simulation runs
parameter_file_path	Path for the Excel file input table containing desired simulation modeling choices (as described in step 2)
size	Number of observations per simulation run
treatment_prevalence	Desired treatment prevalence (between 0 & 1)
treatment coefficient	Desired treatment effect
outcome_prevalence	Desired outcome prevalence (between 0 & 1)
dist	Distribution of the survival time (default= 'E' for exponential; other option 'W' for Weibull)
unmeasured_conf	Identify the unmeasured confounder in the simulations (variable name for unmeasured confounder provided in the parameter_file by the user in quotes e.g "u1")
n_proxies	Number of proxy variables
proxy_type	'binary' or 'continuous'
corr	Desired correlation between unmeasured confounder and proxy variable (between 0 and 1- if more than 1 proxies than independent correlation equal to the value of corr will be generated for all proxy variables)

As described in the manuscript, these simulations generate time-to-event outcomes from a Cox model assuming exponential distribution with a constant baseline hazard by default and has an option to relax this assumption with Weibull distribution. For analysis, 1:1 propensity score matching is used for adjustment and Cox proportional hazard models give hazard ratios.

Step 4: Evaluate results

The function in step 3 will output two tables in a list.

The first table will have average standardized mean differences in the unmeasured confounder and proxies (when n_proxies>0) across simulation iterations at up to 3 levels of adjustment: 1) crude (no adjustment), 2) L1 (adjustment for all the measured confounders), 3) L2 (when n_proxies>0, adjustment for proxies in addition to all measured confounders).

The second table will have average hazard ratios across simulation iterations at up to 3 levels of adjustment: 1) crude (no adjustment), 2) L1 (adjustment for all the measured confounders), 3) L2 (when n_proxies>0, adjustment for proxies in addition to all measured confounders)