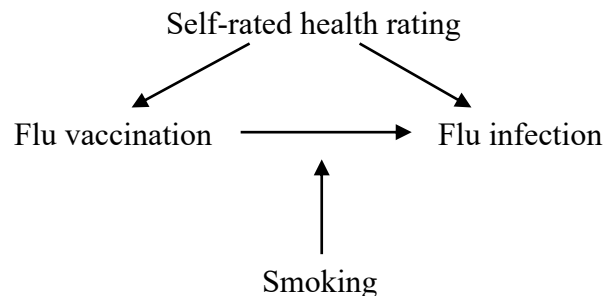Tutorial 2 brief solution

(a) Read the dataset into R. Suppose that based on literature review and a preliminary analysis, the causal structure can be summarized by the following DAG:

Self-rated health rating

Flu vaccination ——————→ Flu infection

Smoking

Fit the above regression model and summarize the results in a table.

Variable	Adjusted odds ratio	95% CI*
Flu vaccination	0.36	0.22-0.57
Self-reported health rating	0.71	0.66-0.76
Current smoker	1.66	1.00-2.66
Flu vaccination × current smoker	2.64	1.01-6.78

*CI: confidence interval

(b) Suppose the same dataset have some missing data ('fluvaccine_m1.csv'). Study the dataset especially on the missingness of the data.

- 10% of the variables included in the final analysis model were missing
- Only 66% of the records are complete cases

(c) It was known that missingness was completely random (MCAR). Carry out a complete case analysis and comparing the results with that in (a).

Variable	Adjusted odds ratio	95% CI
Flu vaccination	0.32	0.17-0.57
Self-reported health rating	0.73	0.67-0.79
Current smoker	2.42	1.38-4.12
Flu vaccination × current smoker	1.81	0.52-5.95

(d) To deal with missing data, multiple imputation method will be used. Create the imputed datasets (say, 50 imputations), based on the variables in the final analysis model. Assess if the variables were imputed sensibly by examining the imputed variables and comparing them with the original dataset.

(e) Perform the regression analysis using multiple imputation method. Compare your results with that in (c).

Variable	Adjusted odds ratio	95% CI
Flu vaccination	0.41	0.32-0.76
Self-reported health rating	0.72	0.70-0.80
Current smoker	2.08	1.20-3.00
Flu vaccination × current smoker	1.42	0.35-3.07

(f) Suppose the same dataset have data not missing completely at random (not MCAR, in 'fluvaccine_m2.csv'). Study the dataset especially on the missingness of the data.

- 10% of the variables included in the final analysis model were missing; more than 20% of the vaccination status were missing
- Only 55% of the records are complete cases
- Subjects with lower self-reported health rating were less likely to report vaccination status

(g) Carry out a complete case analysis and comparing the results with that in (a).

Variable	Adjusted odds ratio	95% CI
Flu vaccination	0.26	0.10-0.55
Self-reported health rating	0.72	0.65-0.80
Current smoker	2.11	1.05-4.00
Flu vaccination × current smoker	1.92	0.34-9.11

(h) Perform the regression analysis using multiple imputation method. Compare your results with those in (c) and (g).

Variable	Adjusted odds ratio	95% CI
Flu vaccination	0.46	0.43-0.84
Self-reported health rating	0.71	0.67-0.76
Current smoker	2.27	1.23-3.15
Flu vaccination × current smoker	0.86	0.27-2.41

(i) Suppose we make use of all available variables to create the imputed datasets. Perform the regression analysis and compare your results with those in (c), (g) and (h).

Variable	Adjusted odds ratio	95% CI
Flu vaccination	0.39	0.20-0.52
Self-reported health rating	0.71	0.68-0.77
Current smoker	2.04	1.18-2.88
Flu vaccination × current smoker	1.67	0.75-5.74

(j) Suppose the same dataset have more missing data (up to 30% missing for key variables, 'fluvaccine_m3.csv'). Study the dataset especially on the missingness of the data.

- 30% of the variables included in the final analysis model were missing
- Only 25% of the records are complete cases
- Subjects with lower self-reported health rating were less likely to report vaccination status

(k) Perform the regression analysis again using multiple imputation method. Compare your results with those in (c) and (h)

Complete case analysis:

Variable	Adjusted odds ratio	95% CI
Flu vaccination	0.23	0.05-0.69
Self-reported health rating	0.68	0.57-0.80
Current smoker	0.63	0.10-2.36
Flu vaccination × current smoker	15.60	1.45-214.27

Multiple imputation:

Variable	Adjusted odds ratio	95% CI
Flu vaccination	0.46	0.24-0.62
Self-reported health rating	0.66	0.63-0.72
Current smoker	1.40	1.09-2.86
Flu vaccination × current smoker	2.58	0.83-6.01