CAnD3 CONSORTIUM ON ANALYTICS FOR DATA-DRIVEN DECISION-MAKING

CRDCN RCCDR Canadian Research Data Centre Network / Réseau canadien des Centres de données de recherche

Research Replicability and Workflow Management

Data Activity

Template program

A human-readable program (your analysis design plan as a researcher) should feature the following elements:

1. Data cleaning
2. (optional) Data merging
3. Definition of analytic sample (e.g. inclusion, exclusion criteria, stratification variables)
4. Table of summary/descriptive statistics (optional: contrasting groups)
5. Table of multivariate analyses

Example from Gibson and Clair (2019), from the researcher program (v2)…

Always go back and revise your program

1. Clean the CCHS
2. Clean the DAD
2. Tag respondents with any of the ACSC conditions in our list
3. Merge the CCHS to the DAD and keep the hospitalizations from the 12 months prior to the CCHS
4. Compare those granting permission to those who didn't consent to the linkage (t)
5. Create a table of summary statistics (t)
6. Create a table of summary statistics for ACSC respondents (t)
7. Compare means of unmet need variable by all available SES metrics (t)

Remember, your program is your writing guide, so keep it current. If you're the type to keep more detailed notes than me, you may want to version the program.

Can you spot a possible meta-efficiency improvement here?

Reproducibility module – Data activity

Version with associated do files (after iteration through "coder" role) (from Gibson & Claire)

1. Clean the CCHS (Part 1 and Part 2) "cchs_clean.do"

   For each year of the CCHS

   i. Drop pregnant women from the CCHS (MAM_037)
   ii. Drop anyone aged over 75 from the CCHS (ANC_Q03)
   iii. Create dummies for marital status, ethnicity, education, income, standardizing across survey waves rolling up where categories are added. (Part 2 data) (MSNC_Q01) (SDC_43A-M) (EDU_04) (INC_03)
   iv. Save a copy to merge for analysis and comparing hospital linked respondents to those with no linkages. "Data_CCHS_cleaned_yy" (where yy is the last two digits of the year of the cchs)

2. Clean the DAD "dad_clean.do"

   For each year of hospital admissions data

   i. Using the list of ICD codes for each set of conditions (ICD-ACSC.xlsx), generate a variable to tag where each appears as a Most-Responsible Diagnosis [waaay too long to list here]
   ii. Generate a list dummy for the two sets of conditions
   iii. Drop hospitalizations where age at discharge is over 75
   iv. Drop observations with no tags
   v. Save a file for each year "Data_dad_cleaned_yy.dta" (where yy is the last two digits of the first calendar year of the discharge data)

3. Merge the CCHS to the DAD and keep the hospitalizations from the 12 months prior to the CCHS "cchs_dad_merge.do"

   For each year of CCHS

   i. Iteratively merge CCHS data to DAD files the year before, the year of, and the year after CCHS date
   ii. After each merge, drop records where survey is taken >12 months from hospital record date (SVY_DATE, ADM_DATE)
   iii. Save a file for analysis