UNIVERSITY of MARYLAND
SCHOOL OF PHARMACY

Pharmaceutical Health Services Research
220 Arch Street, 12th floor
Baltimore, MD 21201
http://www.pharmacy.umaryland.edu

Recommendations for
Working with Large Health Related Data Sets within PHSR

Working with Large Health Data Sets within PHSR:

PHSR is owner of IMS Health Data. PRC is the Data Custodian and will maintain the database secured on PRC's server. Direct questions regarding use and access to the PHSR Vice-Chair for Research and questions about account setup to PRC.

Data characteristics for the 10% sample data set obtained from IQVIA (IMS & Quintiles):

- 12.4 million persons and 17 variables in the population enrollment file
- 74.4 million observations and 3 variables in the benefit coverage file
- 1.2 billion Claims across 13 claims-level files. Each claims-level file has 64 variables

Data Characteristics of the Centers for Medicare & Medicaid Chronic Condition Warehouse (CCW) 5% Sample:

- CCW contains fee-for-service institutional and non-institutional claims, assessment, and enrollment/eligibility data
- Data can be linked by a unique, unidentifiable beneficiary key, allowing research to be conducted across the continuum of care
- Data include 60 pre-defined conditions allowing for easier cohort selection and study of a condition of interest

Best Practices for efficient programming:

Working with very large databases require special attention. In many cases, as the number of observations increase tenfold (e.g., from 10,000 to 100,000), the time to run the same program can easily double. Resources can quickly decrease and degrade SAS processing for all users. Provided guidance is to improve efficiency and operability for all users.

A. *<u>Before writing a SAS program</u>*

- Know the program's purpose. (Steer away from multi-purpose programs if possible.)
- Identify your cohort
- Define required variables
- Outline steps needed to obtain the above
- Create a permanent sample cohort (e.g., 1000 persons)
- Use the sample cohort to pull needed claims.
- Keep needed variables only.
- Test SAS code using the sample cohort and sample claims.
- Insure output is correct before running on the full files.
- Output to permanent files only those datasets needed for downstream processing. <u>Datasets used solely in a program should not be made permanent.</u>
- If you are summarizing claims-level variables to the person-level, DROP all claims-level variables from the output SAS dataset.
- Store only those files needed to complete your research work

B. *Use Programming Code to Aid with File Size*

- Use OPTIONS COMPRESS=YES
  - The more character variables you have in your data set and the longer these text variables are, you can save upward of 80% or more in disk storage. SAS will compress your file ONLY IF the operation will save disk storage. SAS will caution you in the SAS log that it may take longer to read and write a compressed file. We have not seen a significant increase in running time when using the compress option.
- Use the LENGTH statement in your data step for numeric variables whose values are integers.
  - By default, SAS saves numeric variables in 8 bytes. That is 8 characters for each numeric variable. The chart below shows the range of numeric values and the minimum length you can specify for the variable. The LENGTH statement is used on a variable by variable basis.

Byte Length	Maximum Integer
1	NOT ALLOWED
2	NOT ALLOWED
3	8,192
4	2,097,152
5	536,870,912
6	137,438,953,472
7	35,184,372,088,832
8	9,007,199,254,740,990

  - DO NOT USE LENGTH if your numeric variable has non-zero decimal places. The results can be unpredictable.
- When creating flag variables, variables with a yes/no, true/false, present/absent response, think about using a character variable of length 1 (LENGTH yes_no $ 1, for example).
  - Ability to another 2 bytes of disk storage.
  - CAUTION: some statistical procedures in SAS do not support character variables. In order to run procedures using character flag variable:
    - ➢ either recode to numeric variables before running the procedure or
    - ➢ leave them with a length of 3 when creating the data set
- SAS dates should not have a length greater than 4.
  - Example: a date range of 1/1/1800 through 12/31/2030 has SAS date values of -58438 and 25831 respectively.
- Only keep the variables you need. Variables often overlooked are variables used for DO-loops. Unless you are using a variable, you want to keep as the DO-loop index, add the DROP statement for the index variable after the END statement of the DO-loop.
- Obtain foundational skills in programming and in working with large datasets (e.g., over 1 million records). PRC strongly recommends taking/auditing PHSR631 before working with large administrative health/claims data if you do not have prior experience using large databases.
- In order to comply with the Privacy Act of 1974, Health Insurance Portability and Accountability Act (HIPAA) and data use agreement (DUA), data are not to be removed or saved in locations outside of assigned workspace on the PRC server. Do NOT copy source (e.g. IMS) data files to your individual workspace