

Mechanisms for Using the THIN and GPRD Data Resources

Two Mechanisms:

There are two basic mechanisms for working with GPRD and THIN data.

One mechanism relies on the investigator to complete nearly all of the database queries, creation of the data files, and statistical analyses. This mechanism is ideal for investigators with skills in using large databases and statistical software and who may not have sufficient resources to hire biostatisticians and programmers to provide extensive programming. The major advantage of this mechanism is reduced costs. This mechanism is also ideal for pilot studies or studies that require only 10% of the data for sufficient statistical power.

The second mechanism relies heavily on assistance from the Biostatistical Analysis Center within the CCEB to perform all the data management and analysis tasks. This mechanism is preferable for investigators who are inexperienced working with large databases and statistical software. This mechanism is more expensive but may take less time. It is also a useful method for experienced investigators with sufficient financial resources.

Both mechanisms assume that the investigator will define all study variables using appropriate disease, drug, and other codes to create these variables.

It is advisable for investigators who have not used these databases before to meet with representatives of ACARD to discuss which mechanism is most appropriate for their needs. [LINK to Consultation Form and generate automatic email to Rita](#)

Please see Available Funding from CTSA [LINK \(LINK to Funding for Population-Based Database Studies\)](#) for additional information on funding opportunities.

Once the investigator has decided on a mechanism that they will use, it is necessary to apply for access to the databases. [LINK to Online Registration Form](#)

Mechanism A. [LINK to Steps to Accessing and Working with the GPRD and THIN Using 'Victoria'](#)

If data management and analysis are done mostly by the investigator, the investigator starts out testing all aspects of the analysis on a 10% sample of the THIN or GPRD data sets.

- Write and debug any SQL, query, STATA or SAS code to:
 1. Identify the patients (cases and controls; or cohorts) of interest.
 2. Identify the relevant exposure (or treatment) and outcome records for all patients selected in step 1.
 3. Export (to assigned workspace) only those patients and only those data needed for the analysis.
 4. Generate any derived variables needed for the analysis (note that for some types of studies this will need to be done as part of identifying the patients (e.g. calculating age in control group or setting up matching)
 5. Do the analysis
 6. Test all variables, assumptions, and specifications against printouts of a sample of patient profiles, and iterate as needed to generate an analytic data set.
- Generate results
 1. Prepare summary tables of the results.

- 2. Iterate until results are accurate and appropriate to answer the question.
- Write a draft version of the paper with all the required data tables.
- Contact BAC to arrange for creation of a dataset from the full database
 - (contact BAC to make a request at: <http://www.cceb.upenn.edu/pages/apps/ProjectCollab/registerContact.html>).
 (*Note:* BAC will check to make sure that the code they write produces identical results on the 10% sample that is identical to the one on 'Victoria' before running against the full data set. Any discrepancies will be worked out with the investigator.)
- BAC will load the study specific data files from the 100% database onto Victoria
- The investigator will recreate the analytic dataset and rerun the analyses using these files from the 100% database.
- Finalize the paper based on results from the full data set.

Mechanism B. Link to Steps to Accessing and Working with the GPRD and THIN databases With Assistance of BAC

If data management and analysis are done mostly by the BAC, the BAC staff performs the full analyses on the 100% database.

- Contact BAC to make a request at: <http://www.cceb.med.upenn.edu/services/BAC/BAC-contact.php>
- BAC staff writes and debugs any SQL, query, STATA or SAS code to:
 1. Identify the patients (cases and controls; or cohorts) of interest.
 2. Identify the relevant exposure (or treatment) and outcome records for all patients selected in step 1.
 3. Export (to assigned workspace) only those patients and only those data needed for the analysis.
 4. Generate any derived variables needed for the analysis (note that for some types of studies this will need to be done as part of identifying the patients (e.g. calculating age in control group or setting up matching)
 5. Do the analysis
 6. Test all variables, assumptions, and specifications against printouts of a sample of patient profiles, and iterate as needed to generate an analytic data set.
- BAC staff generates results and provides the investigator with computer printouts.
- The investigator prepares summary tables of the results.
- BAC staff, with guidance from the investigator, iterates until results are accurate and appropriate to answer the question.
- The investigator writes the paper with all the required data tables.

Revised April 3, 2007