

## **Ways to Ensure Data Quality of Analysis Results and Output**

**February 8, 2007**

Build up and check your program piece by piece. Check/debug each program module as it is written instead of checking the whole program at the end. Write an analysis/script outline before beginning. Test your program with small datasets (use head and output n number of lines to a different file).

Document your code thoroughly. Keep up-to-date versions in CVS. Create a README.txt file to document precisely what was done (list the input and output files, commands used, etc.).

Explain to someone exactly how data is manipulated in your program to ensure there are no erroneous assumptions or faulty logic. Better yet, have someone go over your code and see if they understand the documentation and what the program does (if they don't then the documentation need work!).

Slow down. Doing things too quickly that turn out to be incorrect only creates more work. Trying to load data into SQL too quickly has pointed out problems with input formats and duplicate values.

Results should be checked to ensure there are no duplicate, missing, or odd-looking values. cut, grep, sort, and uniq can be your friend! Hand calculations can be useful – test your program on simple examples you can do by hand. Compare results with previous analyses if possible.

Run an analyses with different software if possible and compare the results. This has worked well with SAS and justpushplay.

Run analyses twice by two different people if time permits.

Treasure your exceptions: if you have one error, even an apparently minor one, particularly if it is hard to understand, assume that everything else is suspect until the error is identified and corrected.

Assume that you are wrong until proven right rather than the other way around!