Paper 096-2007

Web-based Data Quality Assurance Reports at the National Alzheimer's Coordinating Center (NACC)

Woodrow Deitrich, NACC, University of Washington, Seattle, WA.
Duane Beekly, NACC, University of Washington, Seattle, WA.
William Lee, NACC, University of Washington, Seattle, WA.
Thomas Koepsell, NACC, University of Washington, Seattle, WA.
Walter A. Kukull, NACC, University of Washington, Seattle, WA.

ABSTRACT

A series of approximately 60 tables and reports has been developed for monitoring the quality of data submitted to the Uniform Data Set (UDS) of the National Alzheimers Coordinating Center (NACC). Twenty-nine Alzheimers Disease Centers (ADCs), funded by the National Institute on Aging (NIA), submit data collected on standardized forms from subjects enrolled in the Centers' individual research studies. Data are submitted through NACC's website and error-checked for clear-cut errors before being accepted into the UDS database. Additional quality assurance supervision is done by NACC staff to monitor less clear-cut data inconsistencies, using the Quality Assurance Reports. The QA Reports can be generated on demand by NACC staff via NACC's website, using either all data submitted up to the current time, or any of the past monthly archived data sets. Base SAS and SAS IntrNet are used, including PROC REPORT, PROC TABULATE and ODS. The operating system is Solaris 9.0, but the programs should work on other operating systems. The UDS database is stored in Oracle and all programs are written in SAS. The intended audience should have medium level SAS experience with PROC REPORT, PROC TABULATE, SAS IntrNet and Oracle engine.

INTRODUCTION

The National Alzheimer's Coordinating Center (NACC) receives data on Alzheimer's disease from 29 Alzheimer's Disease Centers (ADCs) located throughout the United States. NACC builds and maintains a database, and makes the data available to researchers at all ADCs, as well as to the National Institute on Aging (NIA), which funds NACC and the ADCs. The purposes of the NACC database are to make widely available a larger set of easily accessible, standardized data than any individual institution could collect, and to foster collaborative research among the ADCs as well as with researchers outside the ADC network.

Data are submitted, error-checked and processed via NACC's website. The database has three separate areas:

1) Working data = data that have been uploaded or data-entered on the website, but have errors or missing data that still require work
2) Current data = the most current data that are complete and have passed the automated error-checks
3) Frozen data = archived data that have received additional scrutiny and have been certified by the Directors of the ADCs that submitted the data

Researchers and administrators at the ADCs and the NIA can access and query the data on the website. NACC staff generate reports from the website, including the Quality Assurance Reports which are the subject of this paper.

The data submitted to NACC are quite complex. Data are collected by the ADCs on nearly two dozen forms. The forms are filled out, usually annually, by different clinicians over a period of time, from the time a patient is enrolled in a research study until death. If an autopsy is performed, a neuropathology form must be submitted as well.

A set of standardized forms, the Uniform Data Set (UDS) is used. The forms are grouped and used in packets, such as the initial visit packet, the follow-up visit packet and the milestones packet. Collecting Alzheimer's data and completing the forms can be challenging. Clinicians work with patients and informants in the field as well as in the clinic. There are requirements as to what constitutes a "completed packet" of forms, which can be controversial at some ADCs. Individual ADCs also have a wide variety of data management strategies for their own databases, which generally contain much additional data not submitted to NACC. They use a variety of processes for submitting the required UDS data to NACC. There are ample opportunities for errors and data discrepancies to arise.

The UDS is a standardized protocol for clinical data collection; its content was determined by the ADC Clinical Task Force. NACC then translated the desired content to data collection forms (electronic and hard copy) and built a relational database to house the data, along with a data submission and access system. Because the protocol, training and specific instructions for data collection are provided directly to clinical staff by NACC, there is a substantial amount of important communication directly between NACC and clinic staff concerning data collection and quality control. NACC's involvement in UDS data collection is not limited to, nor focused on, only communicating with ADC Data Managers or computer systems personnel about how best to submit data files or to complete error checks nor to providing general database advice.

NACC staff consult regularly with the ADCs on data quality issues. A menu of quality assurance programs for use by NACC staff is posted on NACC's website. It allows NACC staff to generate Quality Assurance Reports at any time from any location with internet access, by running programs on the server where the database is housed. The output is returned to the website, where the reports can be reviewed.

SOFTWARE AND HARDWARE

All programs for error-checking, processing data and generating reports are written in SAS v9.13. Base SAS and SAS IntrNet are used. The database is stored in Oracle tables (Oracle 10g). The programs should be compatible with other operating systems as well. NACC uses a Sun Sunfire V880 server running Solaris 9.0.

SUBMISSION OF DATA AND AUTOMATED ERROR-CHECKING

The ADCs submit data to NACC's server via NACC's website, by uploading a data file or by completing web data forms (data entry forms or "real-time" electronic forms). Automated error-checking is launched by the ADC staff who submits the data. Two error-checking processes can be run from the website: simple range checks and alerts.

Range checks are for valid codes (e.g. SEX= 1 or 2), and obvious discrepancies, such as impossible combinations of dates (e.g. a death date that precedes a clinical visit date). Alerts are for suspicious combinations (e.g. a control/normal subject with an extremely low neuropsychological test score), which may nonetheless be legal or correct data. Errors can be corrected online or by uploading corrected data.

Once the data are error-free they can be copied into the current data set, and made available to all the ADCs, the NIA and NACC. At this point NACC can do additional data monitoring for quality assurance.

QUALITY ASSURANCE

The 29 ADCs are independent research entities directed by experts who have different research aims on various medical and psychological aspects of Alzheimer's disease. Because the ADCs have different aims, they may have different recruitment and enrollment protocols. Although the UDS forms are standardized, and a detailed Coding Guidebook of instructions has been written to standardize how the data should be collected, in practice how the forms are used at 29 ADCs may vary. Among NACC's roles are to determine that, as much as possible, all data are being collected, and all data are being collected in the same way, and to facilitate reaching those goals.

Because the ADCs are different, not all error-checking can be standardized and automated. For example, some ADCs may have legitimate reasons, which NACC may be aware of, for some data to be missing, whereas the same data should not be missing at other ADCs. Some ADCs may be enrolling more of a specific kind of patient, which may affect the statistics of the neuropsychological test scores; skewed statistics at one ADC might indicate a data quality problem, while it might not at another ADC.

To incorporate into the automated error-checking all the different types of possible errors and discrepancies, many of which have valid exceptions at certain ADCs, would make the data submission process extremely slow and burdensome for all ADCs. The QA Reports, on the other hand, allow NACC staff to monitor areas like percentages of missing data, and then raise questions with particular ADCs when deemed appropriate.

Therefore NACC has developed a series of approximately 60 tables and other reports which NACC staff use to monitor the quality of the data submitted. Questions which result from these "Quality Assurance Reports" can be addressed to specific ADCs and managed in a more individualized manner, as needed. NACC's statistical staff and other staff develop new topics for QA Reports in-house, and in response to requests from the NIA. Reports can be generated at any time on NACC's website, using either all data submitted up to the current time, or any of the past monthly archived data sets. Some reports access the current data and an archived data set, to show the changes in the database over time.

PROGRAMMING OF THE QUALITY ASSURANCE REPORTS

Most programs use PROC REPORT and/or PROC TABULATE. First, the desired data are extracted from Oracle tables via a LIBNAME statement that allows access to Oracle in a DATA step:

```
libname  mylibref  oracle  user=myoracleuser  password=mypw ;
libname  library  "FRMTS/udsfrmts";
data c1miss;
   format center adcname.;
   *** Get Form C1 test scores from Oracle table ;
   set mylibref.formc1 (keep= center logictst digitfwd digitbwd
                              category trailtst waiststst logic2 bos);
   *** Set flags to 1 for scores missing due to refusal (coded 98 or 998);
   if logictst= 98 then misslog=1;
   else misslog=0;
   if trailtst= 998 then misstr=1;
   else misstr=0;
   (etc.)
   c1count=1;
run;
```

Many reports are tables that show percentages of missing data for each of the ADCs. For example, this table shows the number of neuropsychological test scores that are missing because the patient refused to take the test. If a large percentage of patients are refusing to take tests there will be a lot of missing data weakening the database. However, we cannot make a hard and fast rule about how much missing data is normal or acceptable, and program that rule into the automated error-checking that is done when data are submitted. Some ADCs might be enrolling patients in the late stages of Alzheimer's Disease, due to a particular research aim. At those ADCs, a larger percentage of missing scores might be acceptable, whereas at most Centers it would be cause for investigation. In addition, data may be submitted to NACC and error-checked one patient at a time. A missing score for one patient should not be challenged. Consequently this is best monitored by NACC staff using the QA Reports.

The DATA step can get the test scores from the appropriate Oracle table, then create flags for the test scores that are missing due to patient refusal (see example above). The codes for patient refusal are different depending on the test. PROC REPORT can then count the number of scores that are flagged and compute the percentage of missing scores. The ODS statement is used to generate PROC REPORT output in PDF format, which is then displayed on the website:

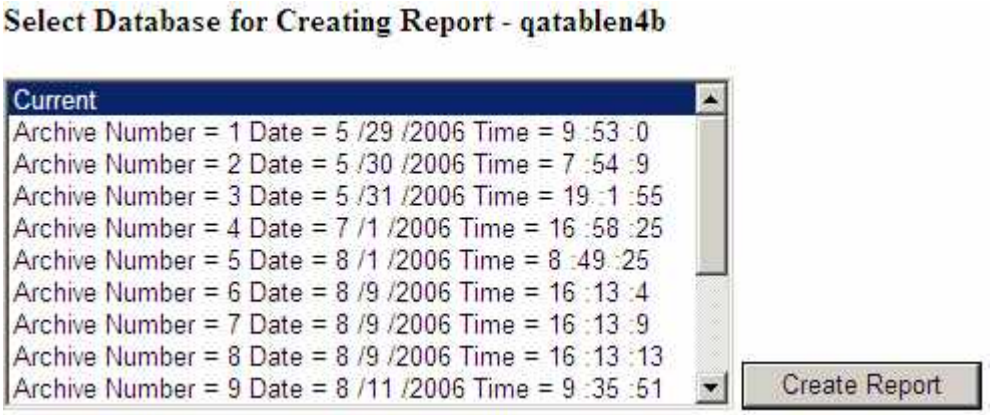
```
ods pdf file = _webout;
PROC report data=c1miss nowindows headline headskip;
      column center c1count misslog logpct missdigf digfpct
             missdigb digbpct misscat catpct misstr trpct
             misswais waispct misslog2 log2pct missbos bospct;
      define center   / group width=17 left "Alzheimer's/Disease/Center";
      define c1count  / sum noprint;
      define misslog  / noprint;
      define logpct   / computed format=3.0 width=7 left
                        'Logical/memory/IA/(%)';
      define missdigf / noprint;
      define digfpct  / computed format=3.0 width=7 left
                        'Digital/Span/Forward/(%)';
      (etc.)

      compute logpct;
         logpct= (misslog.sum / c1count.sum) * 100 ;
      endcomp;
      compute digfpct;
         digfpct= (missdigf.sum / c1count.sum) * 100 ;
      endcomp;

      (etc.)
run;
ods pdf close;
```

The user interface on the website consists of three web pages. The first web page, displayed by a simple HTML file, lists a menu of approximately 60 tables and other reports. Each item on the menu is a hypertext link which runs a separate SAS program on the server. For example, the menu item for the report described above is displayed by this HTML code:

```
<A HREF="/cgi-bin/broker64?_service=uds&TYPEF=FIRST&_debug=0
&_program=dataqual.qatablen4b.sas">
Neuropsych test data missing due to verbal refusal</A><BR>
```

Selecting the above hypertext link in the menu will run the program qatablen4b.sas on NACC's server.
Most of the programs first call a macro in our macros library which displays the second web page. The second page is a simple selection box. It allows the user to select the current data set (see above for definition of "current" data), or any of the archived data sets from the past.

```
%macro archlist;
 libname archive "UDSARCH";
 data _null_;
    set archive.archnums end= eof;
    file _webout;
    if _n_ = 1 then do;
       put '<BR><B>Select Database for Creating Report - ' "&PROG" '</B>' ;
       put '<BR><SELECT NAME=ARCHNUM SIZE = 10 >';
       put '<OPTION VALUE="0" SELECTED>'  "Current";
    end;
    put '<OPTION VALUE="'ARCHNUM'">' "Archive Number = " ARCHNUM " Date = " AMONTH
        "/" ADAY "/" AYEAR " Time = " AHOUR ":" AMINUTE ":" ASEC ;
    if eof then do;
       put '</SELECT> ';
       put '<INPUT TYPE="SUBMIT" VALUE="Create Report" NAME= TYPEP SIZE=300>';
    end;
 run;
 %mend archlist;
```



Once the user selects the desired data set, the program runs and the report is displayed in the third web page.
An example report is Table N4b, which raises questions about data collection at Center C and Center S.

*Table N4b - Percentages of neuropsychological tests missing due to refusal**

Alzheimer's Disease Center	Logical memory IA (%)	Digital Span Forward (%)	Digital Span Backward (%)	Category Fluency (%)	Trail Making Test A (%)	WAIS-R Digit Symbol (%)	Logical Memory IIA (%)	Boston Naming Test (%)
Center A	0	0	0	0	0	0	0	0
Center B	0	0	0	0	0	0	0	0
Center C	20	17	17	18	13	20	18	17
Center D	0	0	0	0	0	0	0	0
Center E	0	0	0	0	0	0	0	0
Center F	0	0	0	0	0	0	0	0
Center G	0	0	0	0	0	0	0	0
Center H	0	0	0	1	0	1	0	0
Center I	0	0	0	0	0	0	0	0
Center J	0	0	0	1	1	1	0	1
Center K	0	0	0	1	1	1	1	1
Center L	0	0	0	0	0	0	0	0
Center M	0	0	0	0	0	0	1	0
Center N	0	0	0	0	0	0	0	2
Center O	0	0	0	0	0	0	0	0
Center P	1	0	0	1	0	1	1	1
Center Q	0	0	0	0	0	0	0	0
Center R	0	0	0	0	0	1	0	1
Center S	9	9	9	9	9	9	9	9
Center T	0	0	0	0	0	1	0	0
Center U	1	0	1	1	0	1	1	1
Center V	0	0	0	0	0	0	0	0
Center W	0	0	0	0	0	0	0	0
Center X	0	0	1	1	1	1	1	1
Center Y	0	0	0	0	0	1	0	0
Center ZA	0	0	0	1	1	1	1	1
Center ZB	0	0	0	0	0	0	0	0
Center ZC	0	0	0	0	1	0	0	0
Center ZD	0	0	0	0	0	0	0	0
Center ZE	0	0	0	0	0	0	0	0
Center ZF	0	0	0	0	0	0	0	0

This table was generated from the UDS Current Database based on 1 /27/2007 data

**** tests coded 98 or 998 on UDS Form C1***

CONCLUSION

Web-based access to the Quality Assurance Reports allows NACC staff to easily work on data quality issues in the office, at home or on the road. The QA Reports are particularly convenient and save time in meetings when a projector connected to the website is used. Current and archived data are available without needing direct access to the server. New programs can easily be added to the existing system to create new reports, as new data concerns and questions arise.

ACKNOWLEDGMENTS

The authors acknowledge the work of the Alzheimer's Disease Centers that submit data to NACC and make constructive suggestions about NACC's website. NACC is funded by NIA grant U01 AG016976.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Woodrow Deitrich
National Alzheimer's Coordinating Center
4311 11th Ave NE, Suite 300
Seattle, WA 98105
Work Phone: 206-543-8637
Fax: 206-616-5927
E-mail: wrich@u.washington.edu
Web: https://www.alz.washington.edu