The Application of SAS® Data Mart in Clinical Laboratory Testing
Cheng Wang, M.D., Ph.D., Jianming He and R Peter Mallon, Ph.D., Quest Diagnostics Inc, Lyndhurst, NJ

ABSTRACT
Business driven data warehouses in the area of clinical laboratory testing have demonstrated preliminary success in providing valuable information to pharmaceutical and managed care organizations (Gilmet et al, 1998). Starting January 1, 1998, a cumulative database has been loaded on daily basis with testing data created by a national leading clinical laboratory. The tremendous of amount of data and resource requirements make it difficult to efficiently perform ad hoc queries or comprehensive analyses without database traffic interruption. A data mart in SAS format has been created and is updated each quarter permitting rapid retrospective analyses while minimizing traffic contention. The SAS Macro codes are presented to demonstrate the effectiveness of SAS data mart.

INTRODUCTION
Quest Diagnostics (www.questdiagnostics.com) is a leading national clinical laboratory currently with more than 30 regional labs and over 1,400 patient service centers. Quest Diagnostics offers more than 2000 clinical tests. On a daily basis, approximately 400,000 requisitions are processed and over 8,000,000 results are reported to physicians. These data are managed by LIS (Laboratory Information Systems) at each regional lab and then transmitted to an information center for uploading into relational database. The database contains 15 tables and more than 150 fields (shown in figure-1). The loading process requires significant time each day, which leaves limited time for data extraction and analysis by company wide users.

This paper describes a resolution of the database access constraints by allocating a SAS data set to perform quick analysis.

DOWNLOAD FIELDS
Major fields from the database were chosen for download to the SAS data mart. These fields include laboratory tests, patient, physician and regional lab information as shown in table-1. Frequently ordered tests are identified by Logical Observation Identifier Names and Codes (LOINC codes) (Forrey et al., 1996). They were put into SAS data set previously as triggers to extract the rest of fields.

Table –1  Major Fields and Formats in SAS Data Mart

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Type</th>
<th>Format</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResultName</td>
<td>Character</td>
<td>$30.</td>
<td>Test result name</td>
</tr>
<tr>
<td>LocalResultCode</td>
<td>Character</td>
<td>$18.</td>
<td>Test result code</td>
</tr>
<tr>
<td>LoincCode</td>
<td>Character</td>
<td>$10.</td>
<td>LOINC code</td>
</tr>
<tr>
<td>Units</td>
<td>Character</td>
<td>$15.</td>
<td>Result unit</td>
</tr>
<tr>
<td>AccessionNumber</td>
<td>Character</td>
<td>$16.</td>
<td>Accession number</td>
</tr>
<tr>
<td>DateOfService</td>
<td>Numeric</td>
<td>Datetime22, 3</td>
<td>Date of service</td>
</tr>
<tr>
<td>NumericValue</td>
<td>Numeric</td>
<td>19.4</td>
<td>Numeric value</td>
</tr>
<tr>
<td>LiteralValue</td>
<td>Character</td>
<td>25.</td>
<td>Literal value</td>
</tr>
<tr>
<td>Age</td>
<td>Numeric</td>
<td>3.</td>
<td>Patient age</td>
</tr>
<tr>
<td>Sex</td>
<td>Character</td>
<td>$1.</td>
<td>Patient sex code</td>
</tr>
<tr>
<td>Zip</td>
<td>Character</td>
<td>$10.</td>
<td>Patient zip code</td>
</tr>
<tr>
<td>AccName</td>
<td>Character</td>
<td>$18.</td>
<td>Account name</td>
</tr>
<tr>
<td>AccNumber</td>
<td>Character</td>
<td>$18.</td>
<td>Account number</td>
</tr>
<tr>
<td>AccZip</td>
<td>Character</td>
<td>$10.</td>
<td>Account zip code</td>
</tr>
<tr>
<td>LabId</td>
<td>Numeric</td>
<td>3.</td>
<td>Lab ID</td>
</tr>
<tr>
<td>LabName</td>
<td>Character</td>
<td>$15.</td>
<td>Lab name</td>
</tr>
</tbody>
</table>

SAS CODES
Data extraction is performed each quarter through the SAS pass through facility. As the data are extracted the standard statistics output is calculated. The SAS codes are as follows.

* SET UP *;
libname data1 "/XXXXX";
libname library "/XXX/library/v8";
filename loinc "/*XXXXX/XXXX*X"; data loinc code;
infile loinc;
input loinc;
proc sql noprint;
create table loinc1 as"
select distinct loinc from loinc1;
select count (loinc) into: obs from loinc1;
select distinct loinc into : list separated by   ' '  from loinc1;
quit;
%global  list;
% EXTRACT DATA FROM SYBASE*;
proc sql;
connect to sybase
(server=XXXXX
database=XXX user=XXX password=XXX);
%put &sqlxrc; %put &sqlxmsg;
%macro pull(loinc);
create table data1.inv&loinc as

DOWNLOAD FIELDS

Figure-1  Database Structure
create table num as
    select *
    from connection to sybase
    (select distinct LoincCode from out1 where n=0)
    order by LoincCode, lab;

create table outb as
    select LoincCode, LabId, count  (literal) as n from data1.inv
    group by LoincCode, LabId;

create table out1b as
    proc sql;
    drop nmiss;
    pctmiss=n/(n+nmiss);
    result=put (LoincCode, $loinc.);
    set out1;
    %let loinc=%scan (&list, &i);
    %do i=1 %to &obs;
        %macro merge;
        %join;
        %end;
    %end;
    %let loinc=%scan (&list, &i);
    %do i=1 %to &obs;
        %macro join;
        %end;
    %mend;
    %pull (&loinc)
    %let loinc=%scan (&list, &i);
    %do i=1 %to &obs;
        %macro join;
        %end;
    %mend;
    and convert ( int,rs.LoincCode) = &loinc
    and rq.DateOfService between  '20001101' and '20001130'
    and rq.AccId=ac.AccId
    and rq.PatientId = pt.PatientId
    and rd.RequisitionId=rq.RequisitionId
    where
        account                    ac
        patient                     pt,
        requisition               rq,
        result-detail            rd,
        result                       rs,
        from
        pt.Sex, pt.Zip,
        rq.AccessionNumber, rq.DateOfService, rq.Age,
        pt.Sex, pt.Zip,
        ac.AccNumber, ac.AccName, pl.AccZip
    from
    result rs,
    result-detail rd,
    requisition rq,
    patient pt,
    account ac
    where
    rs.ResultId=rd.ResultId
    and rd.RequisitionId=rd.RequisitionId
    and rd.PatientId = pt.PatientId
    and rq.AccId=ac.AccId
    and rq.DateOfService between '20001101' and '20001130'
    and convert ( int,rs.LoincCode) = &loinc

* OUTPUT RESULTS *
proc print data=num1;format pctmiss mean median std skewness kurtosis p5 p95 skewness kurtosis;
title1 "Summary statistics";
title2;
data null ;
    set num1 end=_FIEOF;
    %let _FIERR_ = 0;
    %let _FIREC_ = 0;
    %file inventory;
    %let _EFIREC_ = 0;
    %let _EFIERR_ = 0;
    %let _EFIOUT = 0;
    %let _EFIREC_ = 0;
    %let _EFIOUT + 1;
    put loinc code ',' result ',' lab ',' lab-id ',' n ',' pct-miss ',' mean ',' std ',' median ',' mode ',' p5 ',' p95 ',' skewness ',' kurtosis',' end;
    do;
        put loinc code ',' result ',' lab ',' lab-id ',' n ',' pct-miss ',' mean ',' std ',' median ',' mode ',' p5 ',' p95 ',' skewness ',' kurtosis',' end;
        do;
        end;
    if _ERROR_ then call sympit ('_FIERR_');
    if _FIEOF then call sympit ('_FIREC_/ _FIOUT');

* TABLE-2 *
pro csort data=data1.inv;by LoincCode LabId ;
proc univariate data=data1.inv noprint; by LoincCode LabId;
proc sort data=data1.inv; by LoincCode LabId;
proc print data=data1.inv;format pctmiss mean median std skewness kurtosis p5 p95 skewness kurtosis;
title1 "Summary statistics";
title2;
data null ;
    set num1 end=_FIEOF;
    %let _FIERR_ = 0;
    %let _FIREC_ = 0;
    %file inventory;
    %let _EFIREC_ = 0;
    %let _EFIERR_ = 0;
    %let _EFIOUT = 0;
    %let _EFIREC_ = 0;
    %let _EFIOUT + 1;
    put loinc code ',' result ',' lab ',' lab-id ',' n ',' pct-miss ',' mean ',' std ',' median ',' mode ',' p5 ',' p95 ',' skewness ',' kurtosis',' end;
    do;
        put loinc code ',' result ',' lab ',' lab-id ',' n ',' pct-miss ',' mean ',' std ',' median ',' mode ',' p5 ',' p95 ',' skewness ',' kurtosis',' end;
        do;
        end;
    if _ERROR_ then call sympit ('_FIERR_');
    if _FIEOF then call sympit ('_FIREC_/ _FIOUT');

STANDARD STATISTICS
The data have been summarized and aggregated by LOINC codes. A set of statistical parameters has been calculated for each test. Table-2 provides an example
EXAMPLE 1 (OPTIONAL OUTPUT)
Results (e.g., frequency, mean) of a specific test can be grouped according to patient demographics such as age and sex and are presented in figure-2.

EXAMPLE 2 (OPTIONAL OUTPUT)
Weekly aggregated results for LDL-cholesterol test can be graphed to show the trend of test volume and average value (Figure-3).

Figure-2  Patients Demographic Distribution

CONCLUSION
- A relational database was constructed but the system lacked functionality to perform comprehensive statistical analysis in rapid cycle time. Instant access to data source through SAS was not effectively accomplished. However, a SAS data mart approach to perform retrospective analyses offers a solution to database access.
- Data manipulation can be accomplished in SAS environment to reduce the database access time. Uploading output using SAS programs to the database appears to be a more efficient and practical approach.
- The data in SAS data mart exist as raw data, which allows flexible statistical analyses. With the combination of standard statistics and optional programs, more than 70% of the requests of our analysis can be met.

TRADEMARK
SAS is a registered trademark of the SAS Institute Inc, Cary, NC and other countries. Other brand and product names are registered trademarks or trademarks of their respective companies.

REFERENCES
Gilmet GP, Mallon RP, Griffin BT and JJ Lewandowski. The use of an integrated clinical laboratory and pharmacy diabetes database to provide physician performance feedback in an IPA-Model HMO. Journal of Ambulatory Care Management. 1998; 1, 12-23
Quest Diagnostics web site: http://www.questdiagnostics.com/

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the author at:
Cheng Wang
QUEST Diagnostics
1290 Wall Street West, 2nd Floor
Lyndhurst, NJ 07071
201.729.8978 (Direct)
201.729.8998 (Fax)
Wangc@questdiagnostics.com
http://www.questdiagnostics.com/

Jianming He
QUEST Diagnostics
1290 Wall Street West, 2nd Floor
Lyndhurst, NJ 07071
201.729.8979 (Direct)
201.729.8998 (Fax)
Hejian@questdiagnostics.com
http://www.questdiagnostics.com/

R Peter Mallon
QUEST Diagnostics
1290 Wall Street West, 2nd Floor
Lyndhurst, NJ 07071
201.729.8976 (Direct)
201.729.8998 (Fax)
Mallonp@questdiagnostics.com