

Clinical Database acceptance: what statistical review checks are necessary to validate a database?

Genevieve Jehl, Quintiles, Strasbourg, France

ABSTRACT (HEADER 1)

Statistical checks for database acceptance before database lock of a clinical trial are not widely used. Additionally, today, interim analyses or analyses for safety review are more and more common, and hence the teams need to work on unclean or semi-clean data. There is therefore a specific need to know what inconsistencies exist in the data in order to evaluate their effect on the statistical analysis.

Some checks performed by the biostatistics group are necessary to identify major data issues, and help to ensure a clean and complete database for the final analysis after database lock.

INTRODUCTION (HEADER 1)

This paper will present why it is important to involve the biostatistics group in the cleaning process to define edit checks, and why an additional process of statistical review checks is necessary to find missing or inconsistent data.

DATA CLEANING PROCESS (HEADER 2)

The approach of the data is different depending whether you are a data manager or a biostatistician.

The work of the data manager is to get a clean database, all data have usually the same level of cleaning. In most cases, the query "height is missing" and "date of death is missing" is dealt in the same way, and same effort will be applied to get the answer from the investigator. But the second issue may be much more important for the biostatistician, giving a critical information on treatment efficacy/ safety.

It is important for the whole team to discuss and define the list of important data and important checks to focus on.

STATISTICAL CHECKS (HEADER 2)

In addition to data management checks, statistical programmers will program some statistical review checks that will focus on important data (like efficacy or safety data), or missing data (to evaluate the number of missing data points and potentially define imputation methods). Examples of standard statistical checks will be presented.

CALENDAR OF STATISTICAL REVIEW (HEADER 2)

The timing of these checks is important to allow time to ask questions to the investigators and receive the answers without delaying the database lock. Different processes will be described with advantages and disadvantages.

PROCESS OF DATA CLEANING

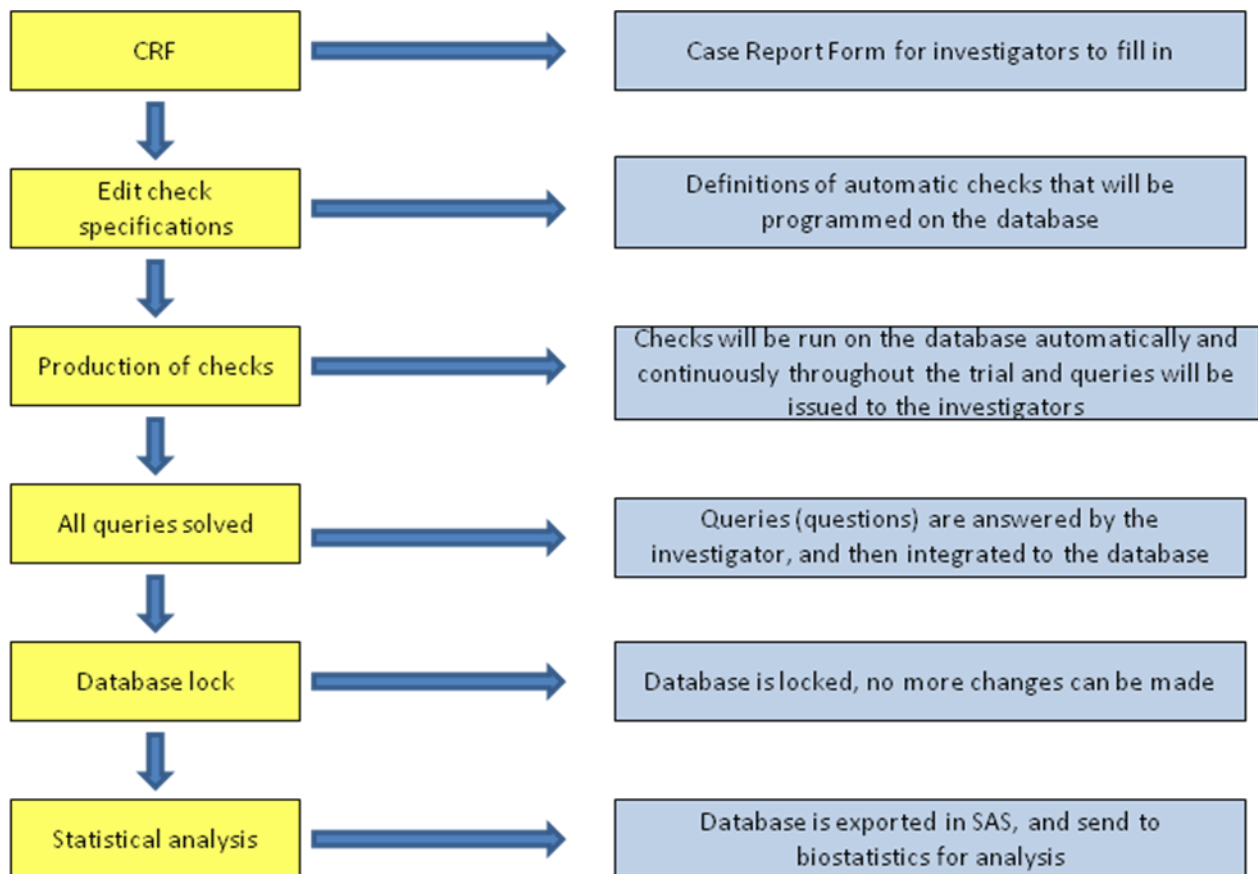
The data cleaning process is defined at the beginning of the project, when defining the structure of the database, the entry guidelines, the data transfer guidelines, as well as the checks that will be applied.

The whole project team needs to be involved in the process, especially the biostatistician to guarantee that the decisions made are in accordance with the protocol and are aligned with the needs of the statistical analysis.

USUAL DATA CLEANING PROCESS (HEADER 2)

The usual cleaning process is:

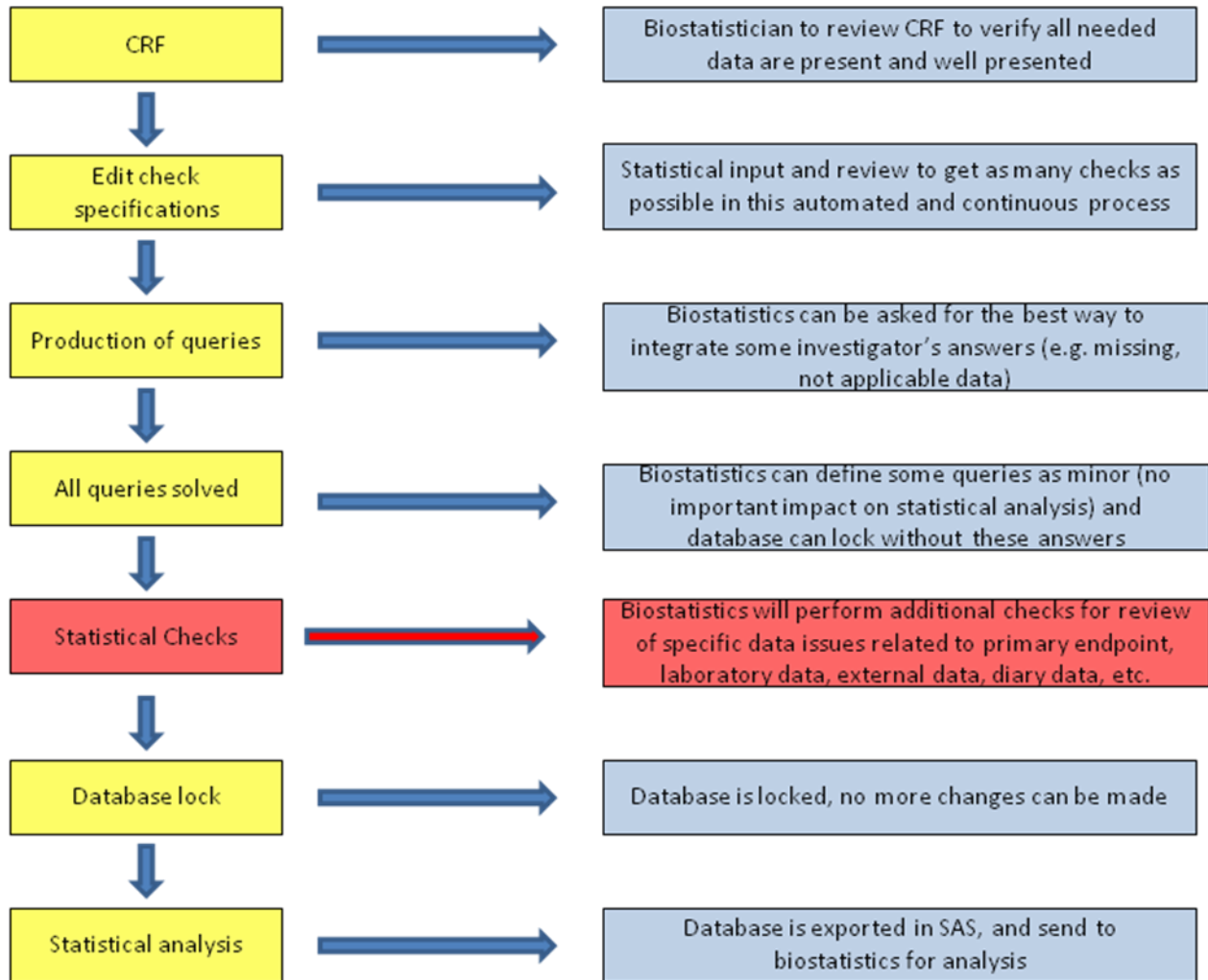
PhUSE 2009



Statistical checks are done in addition to the usual edit checks automatically generated within the database system. They are checked either throughout the study, or at least before database lock, to aid the review of the database and the decision to lock the database.

Biostatistics will appear in this process as:

PhUSE 2009



These statistical checks can be performed after that all queries are solved, to see whether there is any other outstanding data issue that could have an impact on the statistical analysis, or could be done during the course of the trial at the same time as the production of queries, to identify the issues as soon as possible and get feedback from the investigators at the same time.

INTERIM ANALYSIS (HEADER 2)

Outside the usual ongoing process of data cleaning throughout the clinical trial, more and more analyses are performed before the final lock. Interim analyses are planned, especially on phase II trials, as results are needed before the lock to show efficacy of treatment or to stop the trial as soon as possible, for instance because of:

- The increase of costs of the development of a molecule
- The decrease of the success rate for the drugs developed
- More pressure due to generics

There may be two types of analysis during the trial. I will use the term “interim analysis” for any analysis that is done before database lock.

Interim efficacy analyses have been done in the past and are now widely used in phase II trials. The purpose of such analyses could be:

- check the efficacy of the drug with the possibility to stop the study early,
- futility analysis (with stopping rules if the drug does not work)
- sample size re-estimation
- dose selection (stop further recruitment into a dose arm)

PhUSE 2009

Safety analyses are conducted to monitor the safety, mainly on phase II and III trials. Additionally, safety monitoring committees have been highly recommended by authorities to review safety outputs on a regular basis.

These can be referred to in several ways; DSMB (Data Safety Monitoring Committee), DMC (Data Monitoring Committee), IDMC (Independent Data Monitoring Committee) or SMC (Safety Monitoring Committee).

In these cases, it is not always possible to clean the database before each interim analysis. For interim analyses, the database is not locked, we make what we call snapshots. A snapshot is an extract of a database at time t .

A snapshot can be made at any time, but a light review of the database is recommended beforehand. Another possibility is to clean on an ongoing basis but to focus on data up to a specific timepoint (named cut-off date). All data up to that cut-off date will then be cleaned, and although the snapshot is made on the whole database, the statistical analysis will be performed on a restricted set of data in the database export up to the cut-off date.

CLEANING OF AN INTERIM ANALYSIS (HEADER 2)

There are several possibilities to get the optimum level of cleaning of the database for a snapshot, depending on the goal of the analysis to be performed.

Several experiences of ways of dealing with that are:

- The trial team can define specific checks that are deemed important for the interim analysis and the data manager will focus on these, this is then a manual process and can delay the process of data cleaning for the other checks.
- The trial team can review all outstanding queries (not answered by the investigator) one by one a few days before the snapshot during a data review meeting and define the major ones to be solved before the snapshot, this is then again a manual process and can be very time-consuming if there are many outstanding queries to be reviewed.
- Perform statistical checks to identify missing or inconsistent data for the analysis to be done, this is again a manual process and is time consuming but gives a better overview of the impact of data issues on the forthcoming statistical analysis
- Leave the database as it is, and perform the statistical analysis on possibly inconsistent data, this may enable more data to be included in the snapshot but the risk is the need to restart the whole process if the results are not interpretable!

All these processes have advantages and disadvantages, which one is chosen may depend on the applicability of each, on the time available, on the number of queries, on the aim of the analysis etc.

SPECIFIC ISSUES OF EXTERNAL DATA (HEADER 2)

External data are the data coming from a source other than the clinical database. These can be sent to the statistical programmer directly or imported into the database, but in both cases issues may arise with this type of data:

External data are usually randomization data, laboratory data, medical review data, adjudication data, etc.

Early in the process it is key to obtain a dummy file and specifications to agree on the file structure.

In the CRO world, we sometimes get external data with no details on the level of cleaning or the cleaning process used from the third party. This makes our internal cleaning process more difficult, and delays in the database lock can occur.

Regarding external data, especially diary data, checks are not usually applied and the data need to be analyzed as they are. It is therefore highly recommended to involve the biostatistician as much as possible in the definition of process for the collection and review of these external data.

WHY SHOULD THE STATISTICIAN ACCEPT A DATABASE? WHY IS THERE A NEED FOR STATISTICAL CHECKS IN ADDITION TO DATA MANAGEMENT CHECKS? (HEADER 1)

It is important that the biostatistician accepts the database, to be sure that all issues have been checked and discussed. This process will be used for the statistical analysis either to adapt some programs/ analyses, or to include some notes for the reviewers and the clinical study report.

WHY IS THE DATA MANAGEMENT CLEANING PROCESS NOT ENOUGH? (HEADER 2)

Statistical programmers perform some verification tasks after finalization of the query process and before database lock. They will have an in depth look at the occurrence of missing data, outliers, as well as verification of the distribution of the data for the application of the statistical models.

They also check that their programs run without any problems, they check the outputs, they finalize the analysis populations.

PhUSE 2009

Issues can arise at this point if no statistical checks are done and can induce:

- Creation of new queries
- Changes in programs to take into account data issues
- Additional sensitivity analysis with exclusion of extreme values for example
- Change outputs to include notes for the clinical study report
- Etc.

That is why it is important to get a review of the data with a statistical viewpoint before the end of the trial.

In the case of a snapshot of a database for an interim analysis, or a data safety monitoring review, large efforts are made to get all data cleaned. During the cleaning process, recruitment can be stopped, one part of the database can be cleaned and the remaining data will be handled with delays, monitoring visits can be increased to get answers quickly, etc. These have direct impact on costs, thus it is really important to discuss all these issues in advance and define the cleaning process at the start of the trial.

WHAT ABOUT BIG CLINICAL TRIALS? (HEADER 2)

The cleaning process becomes more difficult when more patients, more countries, more data are handled.

In big oncology or cardiovascular trials, it may be difficult to have a database that is completely clean.

At the time of database lock, instead of checking all outstanding queries, it could be better to focus on specific data and base the decision of resolving or not the queries by checking the number of occurrences.

For example, age can be a prognostic factor for efficacy. Some patients have their age missing. If it occurs for 5% of patients, it can be decided to perform the analysis without these 5% of patients.

But if age is missing for 20% of patients, it may be decided that queries need to be answered before database lock.

Statistical Programmers can program with much more flexibility these types of checks, or consistency checks between data from several sources, and biostatisticians will be able to estimate what level of unclean data is acceptable.

WHEN SHOULD THIS STATISTICAL REVIEW TAKE PLACE? (HEADER 2)

The timing of the statistical review needs to be defined at the beginning with the whole team. Usually the checks are done for the data review meeting (before database lock).

This meeting cannot be held too early because we may not see all the issues, and the danger is that a second meeting and a second review needs to be done.

This meeting cannot be too late either because it may delay the database lock by waiting for the last investigator's answers.

The timing of the data review meeting will depend on each trial, on the amount of issues, on the availability of investigators, etc. but a reasonable time would be between 2 weeks and 2 months before the planned database lock.

The statistical review can be done as an ongoing basis during the course of the trial. But the process of running checks and reviewing the outputs is generally time-consuming and may need a lot of resources. In addition to resources, the risk of review of data issues on an ongoing basis may lead to unblind the study treatments. The biostatistician needs to be careful to the summaries performed to maintain the blind of the trial.

WHAT TYPE OF STANDARD STATISTICAL CHECKS COULD BE RUN ON A CLINICAL DATABASE? (HEADER 1)

A task force was formed within the biostatistics team of Quintiles in Strasbourg to define standard statistical checks and programs to be run for database acceptance.

Examples of these checks with corresponding SAS® code will be presented.

Some checks are made to compare different versions of databases, and check the potential differences between the subsequent database exports.

Other checks are more related to the content of the variables, by checking missing or inconsistent information.

VARIABLES ATTRIBUTES (HEADER 2)

PhUSE 2009

It happens that the database is modified during the course of the trial, and the biostatistician may not be aware of that change, or the biostatistician would like to confirm that this change has been made in accordance with the new specifications. The changes can include deletion/addition of datasets, deletion/addition of variables in the dataset, change of attributes of variables, etc. This is important to be checked, in order to adapt the programs that may have been created on a previous version of the database.

The name of the variables can change from one database export to the next, or variables can be added as necessary. This can be checked using SAS® procedure PROC COMPARE.

Sometimes the variables are the same but the attributes may have changed, label, format or length of variables can be modified. The following SAS® code is used to check subsequent versions of database:

```
%*****;
%***      1.4.      All variables with the same type, labels and formats      ***;
%*****;

data DdVarOld DdVarNew;
    set sashelp.Vcolumn(keep=memname name type label format informat length);
    by memname name libname;
        if upcase(libname)("&rawOLD" then output DdVarOld;
        if upcase(libname)("&rawNew" then output DdVarNew;
run;
data checkvar_1_4(keep=memname name diff valO valN);
    merge DdVarOld(in=old rename=(type=typeO label=labelO format=formatO informat=informO
                                length=lengthO ))
          DdVarNew(in=new rename=(type=typeN label=labelN format=formatN informat=informN
                                length=lengthN )) ;
    by memname name;
    length diff valO valN $200.;
    if old and new;
        if typeO ne typeN then do;
            diff='Type';
            valO=typeO;
            valN=typeN;
            output;
        end;
        if labelO ne labelN then do;
            diff='Label';
            valO=labelO;
            valN=labelN;
            output;
        end;
        if formatO ne formatN then do;
            diff='Format';
            valO=formatO;
            valN=formatN;
            output;
        end;
        if informO ne informN then do;
            diff='Informat';
            valO=informO;
            valN=informN;
            output;
        end;
        if lengthO ne lengthN then do;
            diff='Length';
            valO=lengthO;
            valN=lengthN;
            output;
        end;
end;
```

PhUSE 2009

run;

The output looks like:

1.4. All variables with the same type, label, format and informat

There is 2 variables with different attributes:

Dataset	Variable	Attributes	Value in old Database	Value in new Database
ADMI	TOTDOSE	Format	\$4.	
		Informat	\$4.	
		Length	4	8
		Type	char	num
VITS	DIAS	Format	\$3.	
		Informat	\$3.	
		Length	3	8
		Type	char	num

Note: Only variables in both databases are displayed

DATABASE CONTENT (HEADER 2)

In some cases, biostatisticians and statistical programmers may need to compare values in datasets and see what values have been changed. The SAS® procedure PROC COMPARE can be used again, but we have defined another program to have an aggregate overview of the variables modified and their impact on the analysis. Changes are shown by presenting summary statistics of the variables on the previous and new database.

The SAS® code is not presented but the output looks like this:

1.5. All observations are not changed

All compared datasets are stored in Compare_Database folder. Only modifications or new observations are reported.

```
-----
--> The dataset AB will be compared.
-----

* 98 variables / varlist= STUDY DCMNAME DCMSUBNM SUBSETSN DOCNUM INVSITE INV PT ACCESSTS LOGINTS LSTCHGTS
LOCKFLAG CPEVENT DCMDATE DCMTIME REPEATSN ACTEVEVENT SUBEVE
VISIT QUALIFYV QUALIFYQ LAB LABSN LABATYPE LAB_ID ABREACT ASSAYNO ASSAYTP ASSAYRS BARCODE COHORT CMNT INITLS RATIOTP
ABCONC RSLTSP SAMPTP SIGNIFR SITE SUBJNO TPDY TPHR TP
MN TPMO TPWK TITER UNIT VISITDT VISITTM ASSAYNO ASSAYNOQ ASSAYTPC ASSAYTPQ ASSAYRSC ASSAYRSQ BARCODEC BARCODEQ
COHORTC COHORTQ CMNTC CMNTQ INITLSC INITLSQ RATIOPC RATIO
TPQ ABCONCC ABCONCQ RSLTSPC RSLTSPQ SAMPTPC SAMPTPQ SIGNIFRC SIGNIFRQ SITEC SITEQ SUBJNOC SUBJNOQ TPDYC TPDYQ TPHRC
TPHRQ TPMNC TPMNQ TPMOC TPMOQ TPWKQ TPWKQ TITERC TITER
Q UNITC UNITQ VISITDTC VISITDTQ VISITTC VISITTMQ ABREACTC ABREACTQ VISITDTF
```

Variable	Statistic	Statistic for base values	Statistic for comparison values
ABCONC	Number of non-missing value	0	0
	Mean	.	.
	Minimum	.	.
	Maximum	.	.
ACCESSTS	Number of non-missing value	344	344
	Mean	1521720934.7	1521720934.7

PhUSE 2009

	Minimum	1516098529	1516098529
	Maximum	1541333919	1541333919
ACTEVENT	Number of non-missing value	344	344
	Mean	153.53488372	153.53488372
	Minimum	11	11
	Maximum	738	738

UNAVAILABLE EFFICACY DATA (HEADER 2)

Efficacy data are specific to each trial. Checks can be made on the number of missing efficacy data at a specific timepoint or visit, as well as outliers that can be monitored using summary statistics.

Here we have focused on the check of the analysis populations that can be reviewed on each database export, to monitor any issues in protocol violations, etc. during the trial.

It enables one to check whether the primary analysis will be based on the planned number of patients.

3.1. Evaluability of subjects based on the various analysis populations defined

Summary of number of patients by analysis subsets

	Full Analysis Set	Safety Subset	Per Protocol Subset
Number of patients	581	577	503

OUTLIERS (HEADER 2)

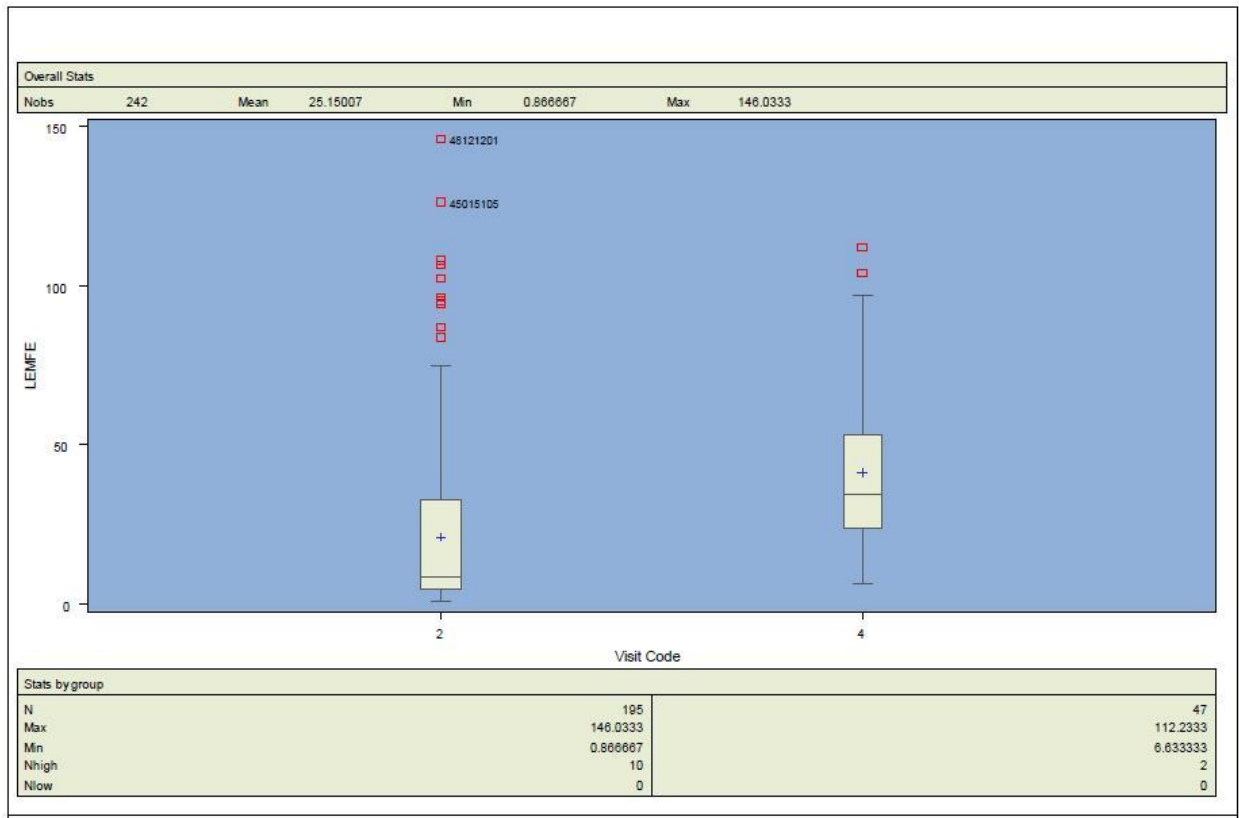
Outliers cannot be easily identified by Data Management, especially when unexpected values occur. An easy way to identify inconsistent data within a numeric variable is to use the box plot presentation.

The following SAS® code can be used:

```
proc boxplot data=plot gout=g;
  plot &var*&boxgr / cframe = vligb
                    cboxes = dagr
                    cboxfill = ywh
                    boxstyle = SCHEMATICIDFAR
                    IDCOLOR =red
                    OUTBOX=OUT_&var
                    name="&var";

  inset nobs mean min max / header = 'Overall Stats'
        cfill = ywh
        pos = tm;
  insetgroup N max min nhigh nlow / header = 'Stats by group' pos=BOTTOM
        cfill = ywh;
  id usubjid;
run;
```

The output looks like this:



Extreme values are clearly identified with this visual check.

CHRONOLOGY OF EVENTS (HEADER 2)

Time windows of dates are not always programmed within data management due to the complexity of the programming, and sometimes software limits.

An easy check is to verify that all visit dates are between the start and end of the study (start and end of the study may need some derivation).

```

%*****
%***          7.1.   Start of study <= visit dates <= end of study          ***
%*****
%*keep the first study date;
proc sort data=&rawNew..enrlment(keep=pt firstDt firstTm);
    by pt firstDt firstTm;
run;
data _71_firstDt;
    set _71_firstDt;
    by pt _firstDt _firstTm;
    if first.pt;
run;

%*keep the last study date;
proc sort data=&rawNew..eos(keep=pt _endDt _endTm);
    by pt _endDt _endTm;
run;
data _71_endDt;
    set _71_endDt;
    by pt _endDt _endTm;
    if last.pt;
run;

```

PhUSE 2009

```

data _7_firstEndDat;
    merge _71_firstDt _71_endDt;
    by pt;
run;

%*check visit dates between the first and the last study date;
%macro check_dates(inData=, DataLabel=);
    proc sort data=&rawNew..&inData(keep=pt dcmdate dcmtime) out=_71_&inData;
        by pt;
    run;
    data _71_&inData;
        set _71_&inData;
        by pt;
        visDat=input(dcmdate,ymmdd8.);
        if dcmtime ne " " then
            visTm=input(compress(substr(dcmtime,1,2)!!'!!substr(dcmtime,3,2)),time5.);
        format visDat date9. visTm time5.;
    run;
    data _checkdata(keep=pt visDat _firstDt _endDt);
        merge _7_firstEndDat _71_&inData;
        by pt;
        if (nmiss(visDat,_firstDt)=0 and visDat<_firstDt) or (visDat>_endDt and _endDt ne .);
    run;
    proc sort data=_checkdata;
        by pt visdat;
    run;

```

The output looks like:

It shows all visit dates that are outside start and end of study.

Patient	Start of study date	End of study date	Visit date
219410001	13MAR2007	26APR2007	28JUN2007
219457002	14AUG2007	05OCT2007	30OCT2007
219457003	28FEB2008	14APR2008	28AUG2008
219583001	02MAY2007	11JUN2007	23AUG2007
219584010	30SEP2008	13NOV2008	07JAN2009
219584011	15OCT2008	04DEC2008	07JAN2009

Note: Start of study date = informed consent date.
DCMDATE used.

LABORATORY DATA (HEADER 2)

Laboratory data are usually a source of problems, due to the different units used, some values that cannot be shown as numeric (e.g. below a limit of detection), some values that seem extreme but may be related to the disease, normal ranges not available, etc.

Several checks can be done:

- If recorded value is given, converted value to be present

PhUSE 2009

- If recorded value is given, recorded unit to be present
- Normal ranges to be present
- Check values versus the normal ranges

Etc.

Here we will check how many values are not converted to SI units.

5.2. If recorded value is not missing then converted value is present

281 patients have CRF results not converted to SI unit:

Laboratory test	Parameters	Number of patients with non converted results
CHEMISTRY	Alanine Amino Transferase	3
	Albumin	13
	Amylase	1
	Aspartate Amino Transferase	2
	Creatinine	3
	Lactate Dehydrogenase	3
	Lipase	1
	Total Bilirubin	13
	Uric Acid	3
	b2 Microglobulin	4
CREAT_CLEAR	Creatinine Clearance (24Hr)	278
HEMATOLOGY	Hematocrit	10
	Hemoglobin	4
	Platelets	4
	Red Blood Cells	4
	Total Neutrophils	4
	White Blood Cells	4

Depending on the number of patients with non converted values, different actions can be taken (get new queries sent to investigators, or change the planned analyses to adapt to these missing data).

CONCLUSION (HEADER 1)

Performing some statistical checks before the database snapshot / lock is recommended to enable any inconsistencies in the data that could induce issues in statistical analysis. It is critical to allow time for this step before database lock to get the database as clean as possible. The biostatistician needs to be involved in the whole process of data cleaning, to provide advice on checks and to take time to review the data on an ongoing basis.

The questions to have at the beginning of the process are: what type of statistical checks is needed on this trial? When do we need to perform these statistical checks? Could it be done just before database lock, or is there a real gain to do this on an ongoing basis, knowing that this review is time-consuming?

The biostatistician needs to stay vigilant to perform checks that will keep the blind of the data. Review of data during the course of the trial is always difficult and the biostatistician needs to be aware of the risk of looking at summaries of data.

Following this, the database should be provided to the biostatistician and statistical programmer in better quality!

CONTACT INFORMATION (HEADER 1)

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

PhUSE 2009

Genevieve Jehl
QUINTILES
Strasbourg
Parc d'Innovation
Rue Jean Dominique Cassini
B. P. 50137
67404 ILLKIRCH-GRAFFENSTADEN
Tel: (+33) 3 88 77 45 38
Fax: (+33) 3 88 77 45 11
Email: genevieve.jehl@quintiles.com

Brand and product names are trademarks of their respective companies.