

Normalized EditChecks Automated Tracking (N.E.A.T.)

A SAS solution to improve clinical data cleaning

Frank Fan, Clinovo, Sunnyvale, CA
Ale Gicqueau, Clinovo, Sunnyvale, CA

WUSS 2010 annual conference
November 2010

Table of Contents

1. ABSTRACT	2
2. INTRODUCTION	3
3. SYSTEM EVOLUTION	5
4. N.E.A.T. BENEFITS	6
5. DESIGN AND IMPLEMENTATION	7
6. N.E.A.T ARCHITECTURE.....	8
7. EDIT CHECKS	9
8. ACTION LIST	12
9. ADDITIONAL MODULES IN N.E.A.T.	16
CONCLUSION.....	17

1. ABSTRACT

One of the major tasks of data management is data cleaning, performed by edit checking. Electronic Data Capture (EDC) Edit checks bring important time and cost savings compared to paper-based systems as they reduce the number of queries requiring interaction between the sites and the sponsor. However, most companies will continue to program Edit Checks in SAS® for different reasons:

- Broader, more complex, and cross-form data issues are difficult to program in EDC
- Many Edit Checks are requested after study roll-out
- Releasing a new EDC with new Edit Checks is costly and complex because of validation requirements

We recommend programming Edit Checks in the initial EDC implementation and manage query resolution within the EDC system:

- EDC systems provide an online query management facility optimized for quick query resolution
- Auto-queries triggered by EDC Edit Checks bring critical savings to the sponsor
- It provides a centralized facility to track and manage all study queries

Normalized EditChecks Automated Tracking (NEAT) is a hybrid solution that leverages the flexibility, agility and power of SAS Edit Checking while keeping the workflow and benefits of EDC query management systems:

- Eliminate redundancies between EDC manual queries and SAS Edit Checks
- Avoid the tracking nightmare of managing SAS queries resolution
- Guarantee high data quality and integrity
- Screen each discrepancy for higher accuracy, clinical relevance and validity
- Provide real-time view for query status for all EDC users
- Identify data issues earlier
- Provide query resolution metrics



2. INTRODUCTION

This paper demonstrates how leveraging SAS/IntrNet capabilities can drastically speed up and improve the data cleaning process in the pharmaceutical and biotechnology industries.

We will present our latest data cleaning tool, N.E.A.T., a data cleaning tool for clinical trials that use the EDC PhaseForward InForm™ system. The same process may be applied for other leading EDC solutions like Medidata Rave, OpenClinica, etc.

The integrated data cleaning system N.E.A.T. is a tool that combines EDC system edit check, EDC integrated query system, and SAS edit checks. All the reports and tools can be presented in a web-based platform developed with SAS IntrNet. Figure 1 is an overview of NEAT.

Encompassed in this paper are the processes for N.E.A.T. system:

- N.E.A.T. Core, which has two main programs: SAS based edit checks and ActionList,
- Site discrepancy email notification,
- Dashboard and internal reporting tools, and
- A web-based platform

N.E.A.T. core consists of two programs and outputs five excel worksheets in two workbooks, ActionList.xls and MasterReport.xls. Besides its core, it may have some more useful tools. The first one is a study overview reporting tool named dashboard, which summarizes and assesses data processing. The second one is called Site Discrepancy Email Notification, which provides related personnel with the discrepancies to process for a specific site. They are both used to fasten the data cleaning process.

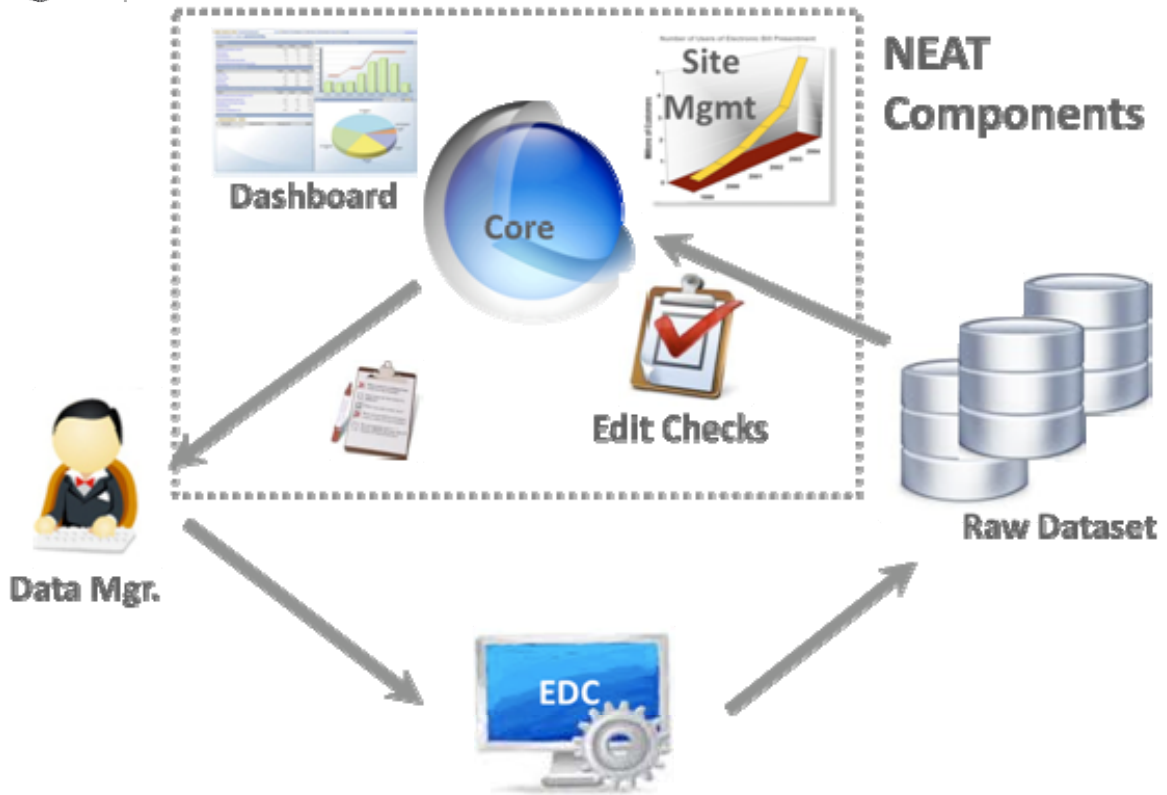


Figure 1. N.E.A.T. Component Structure



3. SYSTEM EVOLUTION

Clinical trials are required to collect and analyze data regarding the safety and efficacy of new drugs and devices. Accurate results are entirely determined by the quality of the data collected.

Without the right tools and the right expertise, data cleaning and query management can prove time-consuming and overwhelming, resulting in trial delays, and considerable loss of revenue. As an example, for a drug like Genentech's Herceptin which generated \$1.2 billion revenues in its first year, each month of delay would have equaled \$106 million of lost revenue.

N.E.A.T. (Normalized Editchecks Automated Tracking) can be used to shorten time for FDA submission by efficiently cleaning collected data as well as easily issuing and tracking queries.

drawbacks of TRADITIONAL Edit Checks solutions

EDC based edit checks are pre-defined at the onset of a study. They are very helpful for data management, since 80% of the data issues are resolved at data entry time.

However, they are unable to tackle complex data issues. EDC only Edit Checks are difficult to program, adapt, and have limitation when dealing with complicated or cross-form checks. Once an EDC system is put to use, modifying or adding more Edit checks is very expensive and time consuming.

SAS Edit Checks with no integration are flexible and able to find all kinds of data issues. They are supplementary means to EDC edit checks. However, they may produce duplicated edit checks and can prove hard to understand.

On the other hand, combining EDC and SAS solution is used to cope with the issues above. But it is proved to become a tracking nightmare and often leads to human errors.



4. N.E.A.T. BENEFITS

Sponsors need additional data quality checks that identify broader data issues other than EDC based edit checks. In this view, N.E.A.T. was designed to efficiently improve the accuracy, completeness and integrity of clinical data.

N.E.A.T. is a cutting-edge SAS-based solution for data cleaning automation that enhances automated edit checks. As an integrated tool, N.E.A.T. combines both EDC and query management in a unique system. It is compatible with any EDC system, in addition to providing Data Managers with faster, more accurate data cleaning and ensuring data consistency.

N.E.A.T. leverages the powerful capabilities of SAS, the indisputable leader in business intelligence. In addition to being compatible with any EDC system, it provides the following benefits:

- Data entry, retrieval, management, and mining
- Report writing and graphics
- Statistical analysis
- Operations research and project management
- Quality improvement
- Applications development
- Data warehousing (extract, transform, load)
- Platform independent and remote computing

It can also be used as a tool to setup communication among data managers, CRAs and CRCs in different sites, thus coordinating all people involved in the data cleaning processes.

- Benefits of N.E.A.T. include but are not limited to:
- Bring together queries from multiple sources
- High standard of data quality and integrity
- Faster and simpler query management process
- Queries resolution
- Automated site performance metrics
- Cross platform, support all EDC systems

5. DESIGN AND IMPLEMENTATION

The system includes two major programs, SAS edit checks and ActionList. The former is used to identify clinical data discrepancies, while the latter is used to conduct integration with already issued queries, previous discrepancies which have been put in exception, or in an EDC query system. This integration also takes into account the discrepancies that have not yet been processed. As the name indicates, it is a list for Data Managers to take actions on. It includes issuing queries and putting queries in an exception list.

To make the system more interactive and flexible, a configuration file is prepared. It contains all the rules to generate discrepancies. Data managers may add more edit check rules and switch any of the rules "ON". Figure 2 shows the workflow and interaction with EDC.

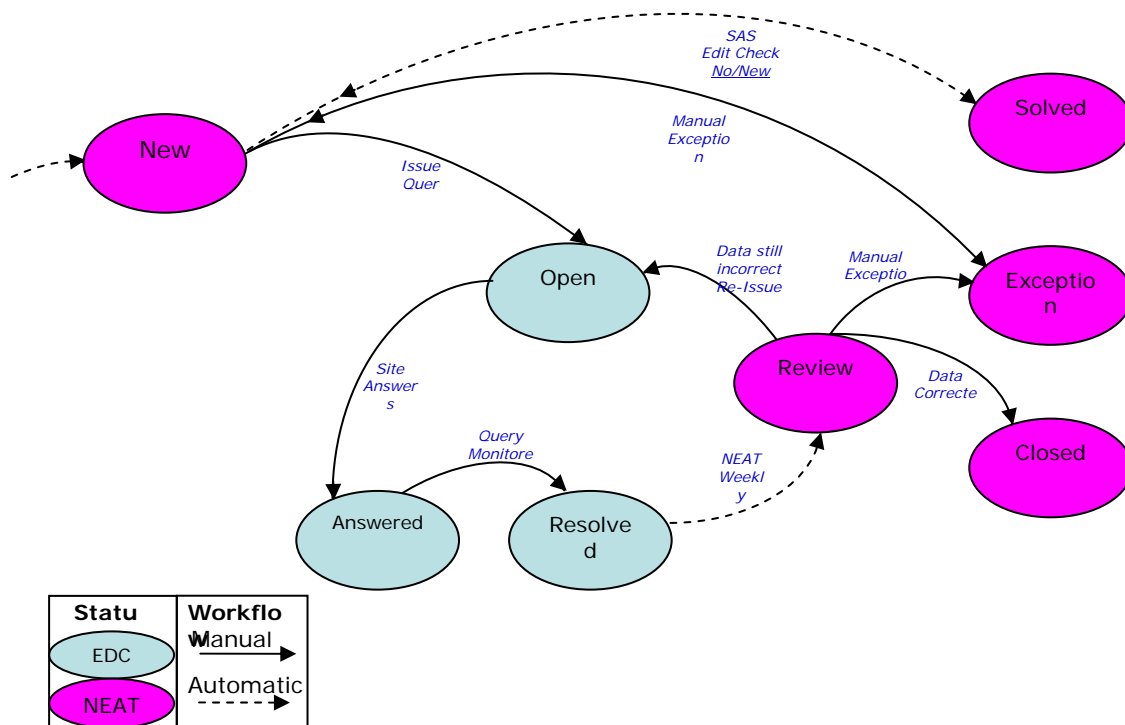


Figure 2. N.E.A.T. Workflow and interaction with EDC



6. N.E.A.T ARCHITECTURE

N.E.A.T. Core realizes these functional parts:

- User configurable mechanism
- Suggested queries to be processed by data managers or clinical monitors
- Yellow forms, i.e. incomplete forms
- Frozen forms with open queries
- Unresolved Closed Queries
- QC New Edit Checks, a new edit checks testing mechanism
- Queries Master View, all SAS and EDC queries with audit trail

By design, many features are configurable and customizable. It offers a flexible and scalable system. For example, we can easily add, modify, update, even delete edit check rules. Data managers can modify the output discrepancy description, turn on/off a specific rules, and add/remove users, without any change to the program code.

7. EDIT CHECKS

Edit check rules are defined by data managers in forms of edit check specification. They are specific to a study and implemented in SAS. The output of the module is a SAS dataset with these columns: "Discrepancy Message", "Data Set", "Alert Date", "Production", "Error Location", "Patient Number".

"Discrepancy Message" is the description which enables data managers to understand the reason for selection.

"Data Set" is an edit check name, also known as a rule.

"Alert Date" is the date when the discrepancy begins to exist.

"Production" takes value of "ON" or "OFF", to branch the discrepancies either to "Suggested_Queries" tab in ActionList or "QC_New_Edit_Check" tab in Master report. This is initially configured by the data manager.

"Error Location" gives the information in which visit and form the discrepancies exist.

"Patient Number" is to identify to which subject the discrepancy belongs.

The edit checks are configurable. Data managers can modify the discrepancy messages and decide if/when to make a check appear. The edit checks are stand-alone and all discrepancies are gathered in one dataset that includes all necessary information.

Here is an example of edit check:

```
* Check 55: Check that the same stents are not entered between the implanted and non
implanted stent forms.;
```

```
/* check between stents 1 and 2*/
```

```
proc sql;
```

```
create table samest12 as
```

```
select a.PATNUM, a.VISITID, a.FORMID, a.ct_recid, a.initial_entry_datetime,
```

```
       a.rrStntRefCode, a.rsStntSerNo, b.rrStntRefCode as code2,
```

```
       b.rsStntSerNo as ref2, a.rsStntSerNo,
```

```
       datepart(a.initial_entry_datetime) as alert_date
```

```
from sdata.p_rfrmSI as a, sdata.p_rfrmSI2 as b
```

```
where a.PATNUM=b.PATNUM and a.rrStntRefCode=b.rrStntRefCode and
```

```
       a.rsStntSerNo=b.rsStntSerNo and a.rrStntRefCode ne " ";
```

```
quit;
```

```
/* check between stents 1 and non implanted*/
```

```
proc sql;
```

```
create table samest13 as
```



```
select a.PATNUM, a.VISITID, a.FORMID, a.ct_recid, a.initial_entry_datetime,
       a.rrStntRefCode,a.rsStntSerNo, b.rrStntRefCode as code2,
       b.rsStntSerNo as ref2, datepart(a.initial_entry_datetime) as
       alert_date
from sdata.p_rfrmSI as a, sdata.p_rfrmnisi as b
where a.PATNUM=b.PATNUM and a.rrStntRefCode=b.rrStntRefCode and
      a.rsStntSerNo=b.rsStntSerNo and a.rrStntRefCode ne "";
quit;

/* check between stents 2 and non implanted*/
proc sql;
  create table samest23 as
  select a.PATNUM, a.VISITID, a.FORMID, a.ct_recid, a.initial_entry_datetime,
         a.rrStntRefCode,a.rsStntSerNo, b.rrStntRefCode as code2,
         b.rsStntSerNo as ref2, datepart(a.initial_entry_datetime) as
         alert_date
  from sdata.p_rfrmSI2 as a, sdata.p_rfrmnisi as b
  where a.PATNUM=b.PATNUM and a.rrStntRefCode=b.rrStntRefCode and
        a.rsStntSerNo=b.rsStntSerNo and a.rrStntRefCode ne "";
quit;

data qc.samestentchk;
  set samest12 samest13 samest23;
  attrib repeat_id length=$20.;
  repeat_id=substr(reverse(trim(CT_RECID)),1,8);
  alert_date=datepart(initial_entry_datetime);
  drop _initialentry_datetime ct_recid;
run;
```

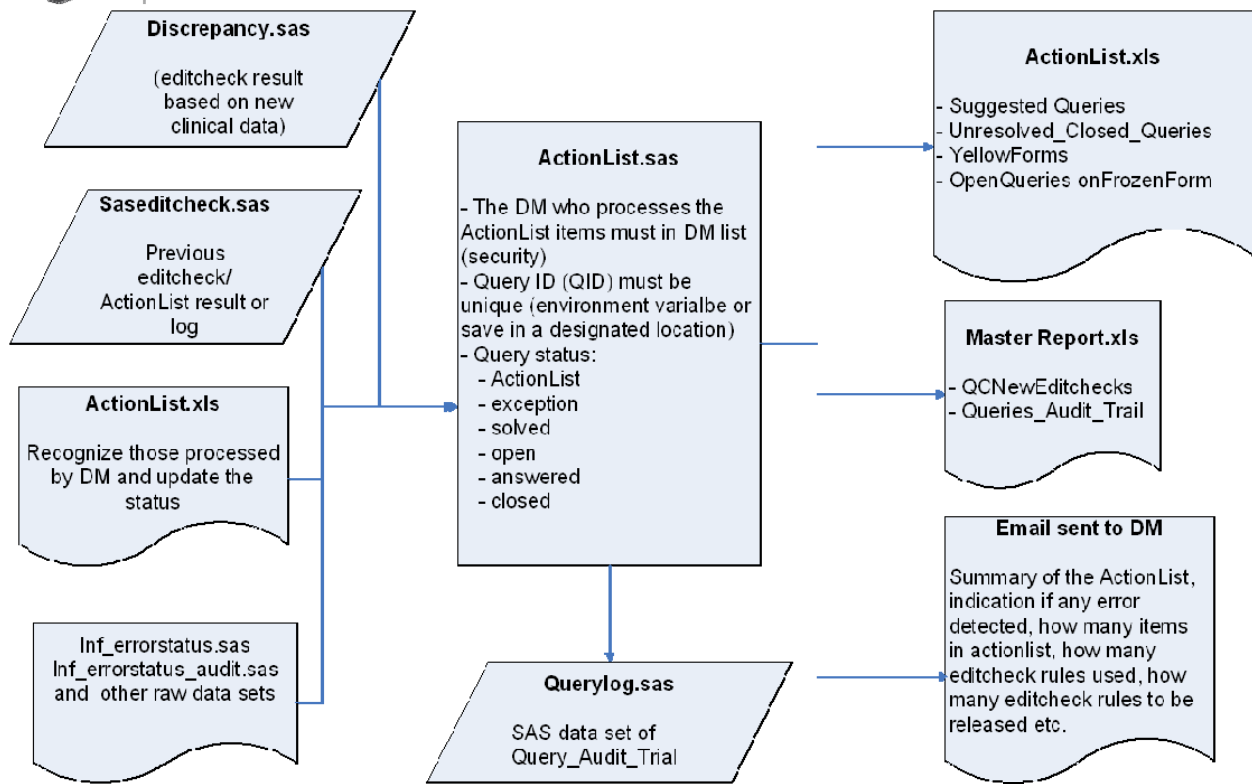


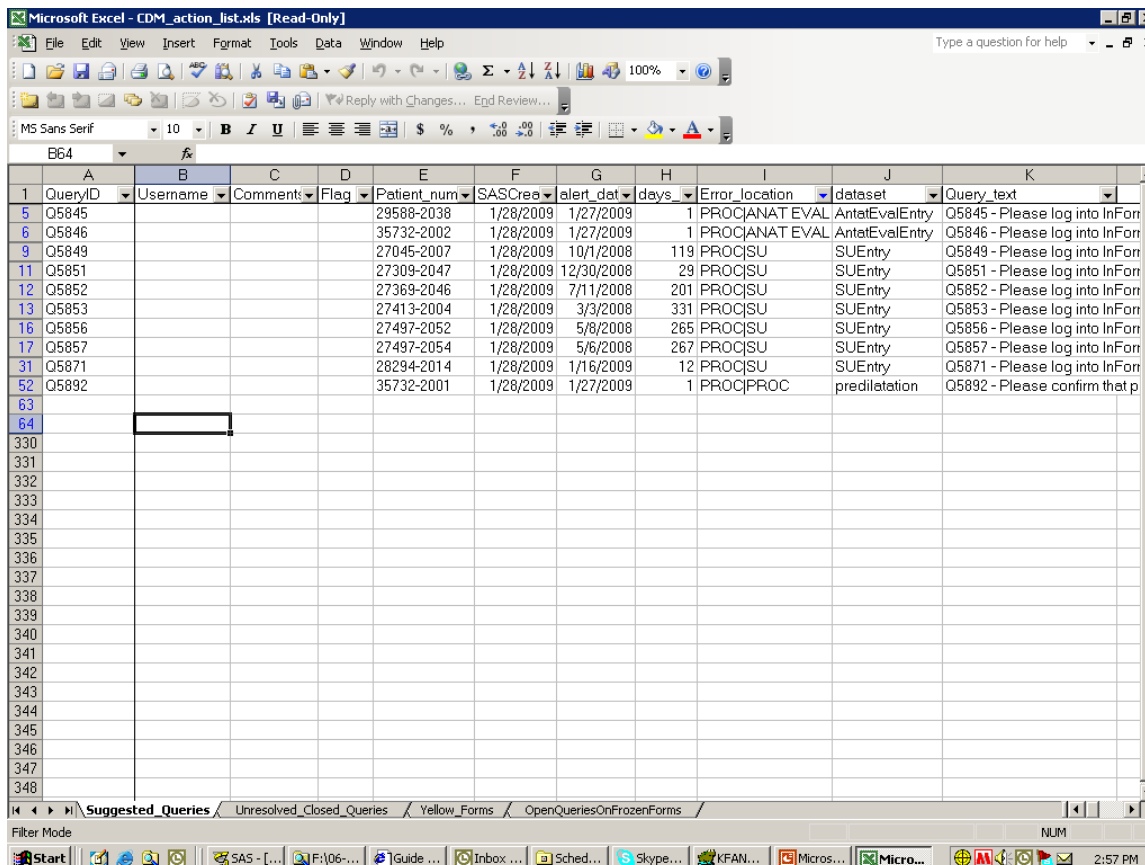
Figure 3. N.E.A.T. core – Actionlist design

8. ACTION LIST

A crucial output of N.E.A.T. is an excel file, called ActionList.xls. It is available for data managers to review daily. Another Excel file output is called MasterReport.xls, which lists new discrepancies uncovered by edit check rules (configured to “OFF”) and all the pending queries. ActionList is also configurable by data managers. Figure 4 shows how the actionlist.sas works.

ActionList is a program with the following inputs:

- Discrepancy.sas, which includes all discrepancies identified by the result of edit check programs
- Saseditcheck.sas, which is the accumulative SAS recognized discrepancies and queries
- ActionList.xls, which is a read-only file that provides information about the discrepancies which have been processed (either a query has been issued or marked as an exception) and the user name information.
- Query information from EDC system



QueryID	Username	Comment	Flag	Patient_num	SASCreated	alert_dat	days	Error_location	dataset	Query_text
5	Q5845			29588-2038	1/28/2009	1/27/2009	1	PROCANAT EVAL	AntatEvalEntry	Q5845 - Please log into InFor
6	Q5846			35732-2002	1/28/2009	1/27/2009	1	PROCANAT EVAL	AntatEvalEntry	Q5846 - Please log into InFor
9	Q5849			27045-2007	1/28/2009	10/1/2008	119	PROCSU	SUEntry	Q5849 - Please log into InFor
11	Q5851			27309-2047	1/28/2009	12/30/2008	29	PROCSU	SUEntry	Q5851 - Please log into InFor
12	Q5852			27369-2046	1/28/2009	7/11/2008	201	PROCSU	SUEntry	Q5852 - Please log into InFor
13	Q5853			27413-2004	1/28/2009	3/3/2008	331	PROCSU	SUEntry	Q5853 - Please log into InFor
16	Q5856			27497-2052	1/28/2009	5/8/2008	265	PROCSU	SUEntry	Q5856 - Please log into InFor
17	Q5857			27497-2054	1/28/2009	5/6/2008	267	PROCSU	SUEntry	Q5857 - Please log into InFor
31	Q5871			28294-2014	1/28/2009	1/16/2009	12	PROCSU	SUEntry	Q5871 - Please log into InFor
52	Q5892			35732-2001	1/28/2009	1/27/2009	1	PROCIPROC	predilatation	Q5892 - Please confirm that p

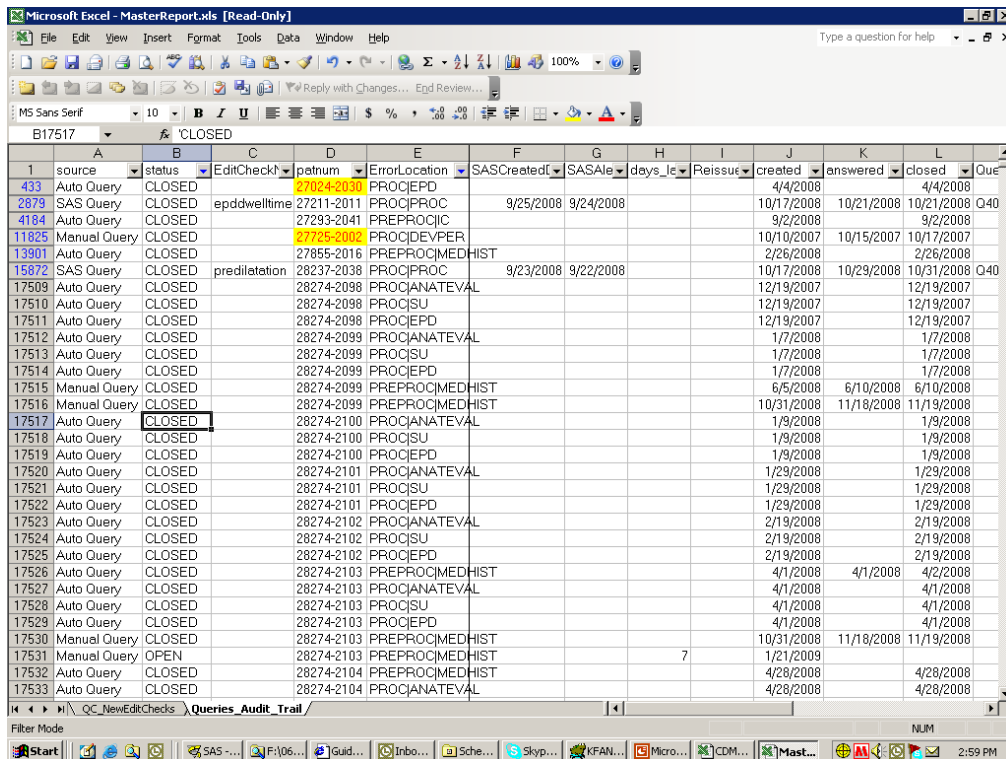
Figure 4. N.E.A.T. core – Actionlist output (suggested queries)

The program generates the below outputs:

- Suggested queries, the most important output to data managers to clean up data
- QCNewEditchecks, discrepancies generated by those edit check rules which are not yet switched to “Production”.
- Queries_Audit_Trail, lists all kinds of queries (auto-queries, manual-queries, and sas-queries) with status
- Email notification. An email is sent to data managers to inform them if the previous actionlist items processed properly, how many edit checks rules remain in production and how many are yet to release
- Some supplementary output files, such as “YellowForms”, “OpenQueriesOnFrozenForm”, and “Unresolved_Closed_Queries” provide data managers with more useful information

The program integrates all inputs, updates the necessary files, and generates the output files. It summarizes all the EDC queries in one dataset and updates the SAS query status accordingly. On the other hand, discrepancies need to be compared with previous identified and processed discrepancies that have the status “updated”. In case an exception already exists, the new discrepancies will not appear in the actionlist output.

Suggested_Queries is a worksheet in one of the outputs of ActionList.xls. It includes the new discrepancies uncovered by the edit checks program, but have those processed in ActionList, listed in ActionList, and already issued queries in EDC before removed. Figure 4 shows the Suggested_Queries output.



1	source	status	EditCheck	pnum	ErrorLocation	SASCreated	SASAle	days_Ic	Reissuc	created	answered	closed	Que
433	Auto Query	CLOSED		27824-2030	PROCIEPD					4/4/2008		4/4/2008	
2879	SAS Query	CLOSED	epddwelltime	27211-2011	PROCPROC	9/25/2008	9/24/2008			10/17/2008	10/21/2008	10/21/2008	Q40
4184	Auto Query	CLOSED		27293-2041	PREPROCIC					9/2/2008		9/2/2008	
11825	Manual Query	CLOSED		27725-2002	PROCIDEVPER					10/10/2007	10/15/2007	10/17/2007	
13901	Auto Query	CLOSED		27855-2016	PREPROCIEMHIST					2/26/2008		2/26/2008	
15872	SAS Query	CLOSED	predilatation	28237-2038	PROCPROC	9/23/2008	9/22/2008			10/17/2008	10/29/2008	10/31/2008	Q40
17509	Auto Query	CLOSED		28274-2098	PROCIANATEVAL					12/19/2007		12/19/2007	
17510	Auto Query	CLOSED		28274-2098	PROCIJSU					12/19/2007		12/19/2007	
17511	Auto Query	CLOSED		28274-2098	PROCIJPD					12/19/2007		12/19/2007	
17512	Auto Query	CLOSED		28274-2099	PROCIANATEVAL					1/7/2008		1/7/2008	
17513	Auto Query	CLOSED		28274-2099	PROCIJSU					1/7/2008		1/7/2008	
17514	Auto Query	CLOSED		28274-2099	PROCIJPD					1/7/2008		1/7/2008	
17515	Manual Query	CLOSED		28274-2099	PREPROCIEMHIST					6/5/2008	6/10/2008	6/10/2008	
17516	Manual Query	CLOSED		28274-2099	PREPROCIEMHIST					10/31/2008	11/18/2008	11/19/2008	
17517	Auto Query	CLOSED		28274-2100	PROCIANATEVAL					1/9/2008		1/9/2008	
17518	Auto Query	CLOSED		28274-2100	PROCIJSU					1/9/2008		1/9/2008	
17519	Auto Query	CLOSED		28274-2100	PROCIJPD					1/9/2008		1/9/2008	
17520	Auto Query	CLOSED		28274-2101	PROCIANATEVAL					1/29/2008		1/29/2008	
17521	Auto Query	CLOSED		28274-2101	PROCIJSU					1/29/2008		1/29/2008	
17522	Auto Query	CLOSED		28274-2101	PROCIJPD					1/29/2008		1/29/2008	
17523	Auto Query	CLOSED		28274-2102	PROCIANATEVAL					2/19/2008		2/19/2008	
17524	Auto Query	CLOSED		28274-2102	PROCIJSU					2/19/2008		2/19/2008	
17525	Auto Query	CLOSED		28274-2102	PROCIJPD					2/19/2008		2/19/2008	
17526	Auto Query	CLOSED		28274-2103	PREPROCIEMHIST					4/1/2008	4/1/2008	4/2/2008	
17527	Auto Query	CLOSED		28274-2103	PROCIANATEVAL					4/1/2008		4/1/2008	
17528	Auto Query	CLOSED		28274-2103	PROCIJSU					4/1/2008		4/1/2008	
17529	Auto Query	CLOSED		28274-2103	PROCIJPD					4/1/2008		4/1/2008	
17530	Manual Query	CLOSED		28274-2103	PREPROCIEMHIST					10/31/2008	11/18/2008	11/19/2008	
17531	Manual Query	OPEN		28274-2103	PREPROCIEMHIST			7		1/21/2009			
17532	Auto Query	CLOSED		28274-2104	PREPROCIEMHIST					4/28/2008		4/28/2008	
17533	Auto Query	CLOSED		28274-2104	PROCIANATEVAL					4/28/2008		4/28/2008	

Figure 5. N.E.A.T. core – Actionlist output (Master report)

Queries_Audit_Trail is a worksheet in an output file, MasterReport.xls, as shown in Figure 5. It compiles all discrepancies for the study, including those:

- Identified by SAS edit check and manually inputted in EDC system
- Identified by SAS edit check but not inputted in EDC system (exceptions)
- Identified by SAS edit check but not processed (not manually input in EDC system and not marked as an exception)
- Not identified by SAS edit checks and manually input in EDC system
- Auto-generated by EDC system (not identified by SAS edit checks, not manually input in EDC system)

Query ID must be unique. Hence a system environment variable may be used to keep the latest value. Each time the program is running, the new entry will assign a value based on that environment variable. To avoid duplicate Query IDs, the discrepancies are recognized using an ID, which includes the edit check data set name, the subject number, error location, i.e. the visit and form of the discrepancies, alert date, and unique ID if necessary. The format is "datasetname_patnum_visitid_formid_alertdate_uniqueid".

Below is the code to generate Queries_Audit_Trail.

```

data querylog;
    attrib status length=$12;
    merge queries(in=a) sdata.SASEditChecks(in=b);
    by QueryID;
    if b and (^a or SASstatus="EXCEPTION") then status = SASstatus;
    if b then source = "SAS Query";
    else if a then ErrorLocation=compress(visit||"||"||form);
    if created ne . and status = "OPEN" then days_late=today()-created;
    drop form visit SASstatus;
run;

proc sort data=querylog;
by QueryID;
run;

```

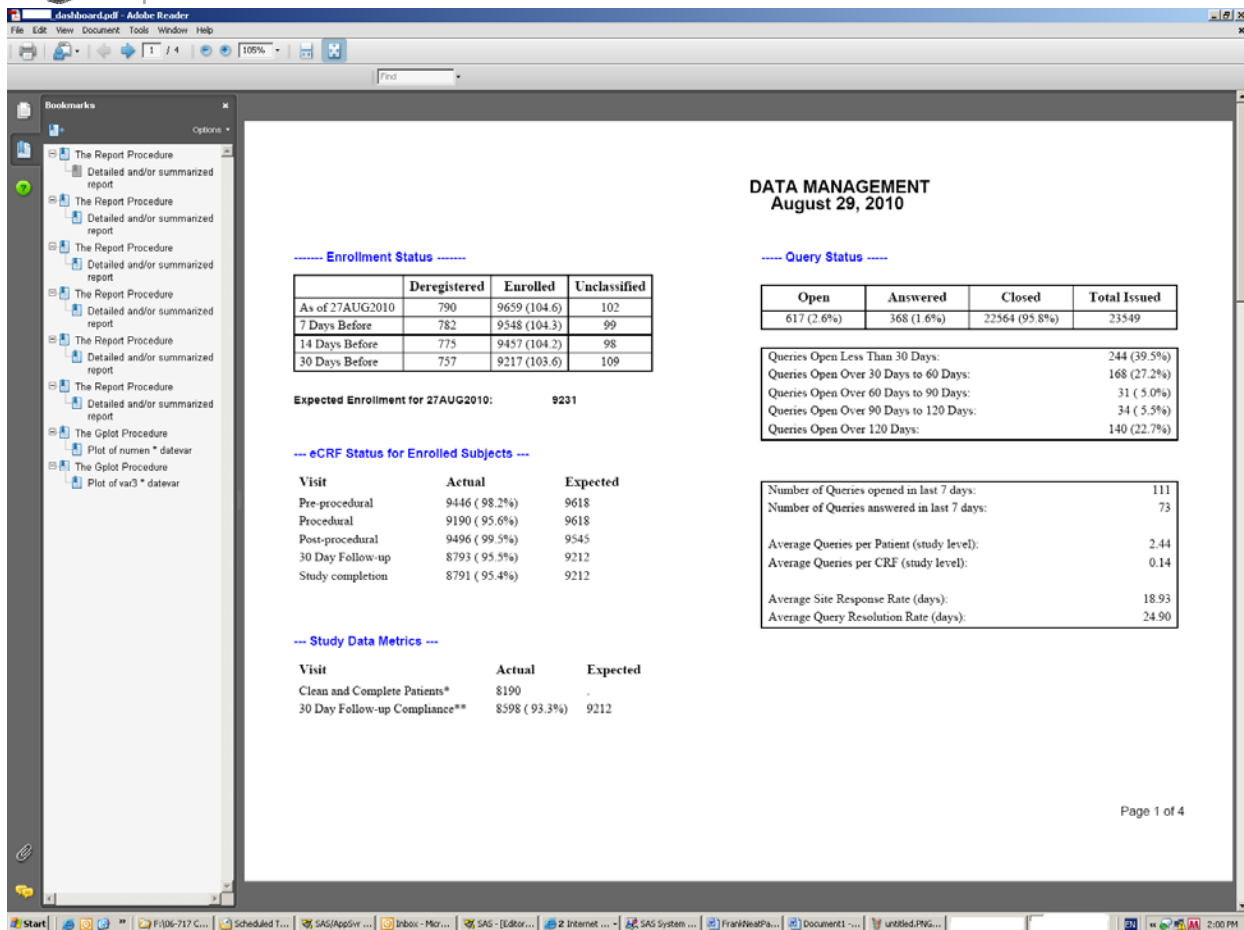


Figure 6. Metrics report – dashboard

9. ADDITIONAL MODULES IN N.E.A.T.

Additional tools were added to help speed up the data cleaning process, such as the dashboard report, site discrepancy email notification, and web platform.

The dashboard report provides users with an overview of the study. The query status on this dashboard lists the number of open/answered/closed queries for different time periods. It is a study level report that gives data managers, project managers, and clinical monitors useful data cleaning information. Dashboard can be used to compare metrics among different studies. Figure 6 is an example of Dashboard report.

Site discrepancy email notification, on the other hand, is a site level report that provides the CRC, the data manager, and the CRA with the outstanding queries and related personnel. This aims to fasten the process of answering, closing queries and correcting clinical data.

Site email notification helps identify data issues earlier, minimize backlogs, support analysis activities, and allow for sooner interpretation and data reporting. With it, CRCs do not need to log in; they are directly aware of data issues, and the consolidated “query action list” enables them to better manager the data cleaning process.

We have extended N.E.A.T. to SAS/IntrNet, a web-based module of SAS, to provide sponsors with a web-based platform. With this enhancement, data processing becomes centralized, interactive, secure, scalable, and user friendly. Benefits also include real time 24/7 access, a lower total cost of operation, and simultaneous access. Figure 7 is a demo of NEAT online.

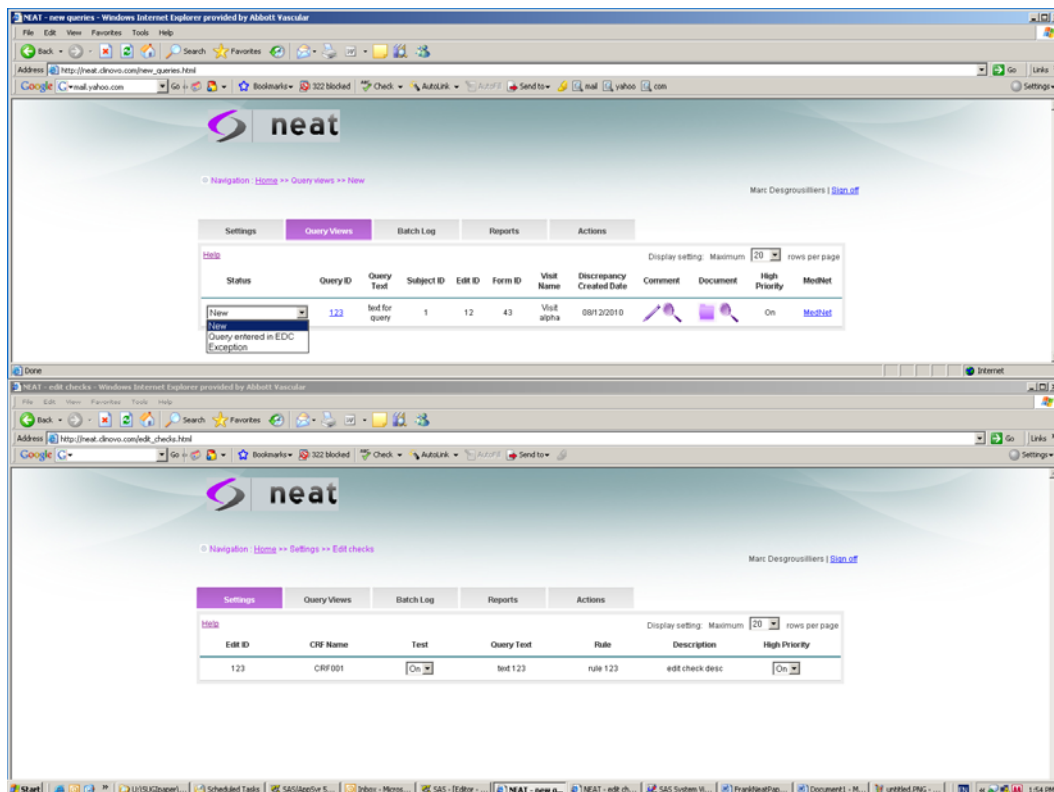


Figure 7. N.E.A.T. web-based interface



CONCLUSION

When EDC appeared a few years ago, many observers thought that it would speed-up data cleaning while improving data quality. Though EDC promises held true on many levels, creating new Edit Checks became more challenging due to the validation requirements for 21 CFR part 11 compliance.

As a result, many companies are undergoing a query tracking nightmare. As more and more companies are using EDC while still relying on SAS for data analysis, a clinical data tool such as N.E.A.T. will become a must-have application to reconcile EDC and SAS queries.



ACKNOWLEDGMENTS

We would like to thank all the people involved in the development of N.E.A.T.: Marc, Jo, Hendra, Romain, as well as the users for their valuable feedbacks.

CONTACT INFORMATION

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

Frank Fan

Application Programmer at Clinovo

1208 E. Arques Avenue, suite 114

Sunnyvale, CA 94085

E-mail: frank@clinovo.com

Ale Gicqueau

President and CEO at Clinovo

1208 E. Arques Avenue, suite 114

Sunnyvale, CA 94085

E-mail: ale@clinovo.com

Phone: +1 800 987 6007

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.