

## How a Clinical Data Repository can support Pharmacovigilance

Pantaleo Nacci, GSK Vaccines S.r.l., Siena, Italy

### ABSTRACT

The adventure of the Novartis (now GSK) Vaccines' Clinical Data Repository continues: clinical trial data in CDISC SDTM format are now available for more than 230 (non-flu) studies, and a new phase has started. In my current role I have a vested interest in finding new or better ways of exploiting this huge mass of data, to keep my clients happy and at the same time contribute to deliver safe and effective vaccines to the market as fast as possible.

### INTRODUCTION

This is my latest paper on the Clinical Data Repository (CDR) initiative, which started soon after the 2009 flu pandemic was over. During that period we received from regulatory bodies all over the world a series of urgent requests for data on clinical data experience with a proprietary adjuvant in all age groups: unfortunately the only vaccine on the market using it was registered in elderly people only, so we were able to answer only partially and even then with some delay.

Now that CDR is available we are able to perform the same MedDRA® searches in the Pharmacovigilance (PV) and clinical trial databases, e.g., allowing the PV colleagues to enlarge the scope of their investigations from just the cases recorded in the PV database to all the adverse event data collected in our trials.

We have also started looking at data visualization packages, e.g., to explore the data graphically, identify safety signal and test new detection algorithms, as well as easily create patient profiles.

### CDR - A QUICK RECAP

Clinical Data Repository is the name of the integrated environment for storing, managing and reporting clinical trial data (and metadata) based on SAS® Drug Development (SDD) version 3.5.

CDR has been developed to revolutionize our ability to:

- Address complex health authority questions quickly and completely
- Produce CDISC-compliant submissions
- Review all available safety data in real-time
- Mine our overall database for scientific and commercial queries
- Improve overall productivity in Clinical Research & Development

For a detailed description of data oceans and how exposure information is stored in CDR, please refer to the paper I presented last year in London, which can be found in the PhUSE wiki.

### THE 'CLASSICAL' PHARMACOVIGILANCE APPROACH

This is the definition of pharmacovigilance found on the EU website: "All medicinal products in the EU are subject to a strict testing and assessment of their quality, efficacy and safety before being authorised. Once placed on the market they continue to be monitored so to assure that any aspect which could impact the safety profile of a medicine is detected and assessed and that necessary measures are taken. This monitoring is called pharmacovigilance."

This has been historically managed by means of a database (in our case Oracle® Argus) containing serious adverse reactions, organized in 'cases'. MedDRA-coded data in these databases are kept homogeneous by always using the latest dictionary version for all entries.

As far as I know analysis of clinical trial data is mostly performed at the single study level, or at most on *ad hoc* poolings of selected studies, often never reused again.

### A SHORT LOOK AT ARGUS QUERIES

Argus includes by default several tools helping with searches on medical terms, including the ability to use all levels of MedDRA terminology up to Standardised MedDRA Queries (SMQs), which are validated, standard sets of MedDRA terms.

Over time we have developed internally a standard form, which the medical doctors in the PV department can fill in to fully specify the search scope and criteria: the goal is to ensure that the implementation path is clear for the operators and final results are as expected.

A ticket is opened in a tracking system with the completed form attached, and the resulting output is added to the same ticket before closure, so that all requests are fully logged.

### ANALYSIS SCENARIOS WITH CDR

As with all such complex systems, CDR was initially just a resource black hole: plenty of work by many people went into setting it up, defining and documenting its processes, etc. but very little came out of it.

## PhUSE 2015

Things started changing once the first lot of SDTM-remapped legacy data, including 153 studies from 1992 onwards, was finally made available. At that time no product had all its data in CDR yet, but in some cases more than 90% was there.

The first tests started by replicating previous calculations used for the yearly updates of Investigator Brochures (IBs): study-level exposure numbers, which in the past had to be derived one by one going through the original CSRs, were obtained by a single PROC FREQ, and the few discrepancies investigated and quickly solved. Of note, in the case of extension studies we were able to easily identify subjects who had been already exposed to the vaccine of interest in the parent study (by using USUBJID as the key variable), thus avoiding double counting them.

In addition to that, the creation of the list of studies where administration of a certain vaccine component/formulation was planned was deemed as useful.

The next step was to identify (all studies including) subjects who have reported at least one adverse event (AE) which was coded to one or more MedDRA Preferred Terms (PTs): once that was done several different scenarios were available, e.g., all AEs reported in selected studies, only AEs reported by subjects exposed to a certain vaccine component/formulation, etc.

To reach these goals two programs were prepared, and several new SAS macros were developed along the way, in addition to those already available as part of our Standard Reporting Software suite.

### CALCULATING PLANNED AND ACTUAL EXPOSURE TO A VACCINE COMPONENT/FORMULATION

This is the high level structure of the first program, calculating both planned and actual exposure to a certain vaccine component or formulation:

1. Call macro %MRG\_OCEANS to create overall TE domain from complete and ongoing oceans
2. Identify ongoing, blinded studies using the TS domain from ongoing ocean
3. Call again macro %MRG\_OCEANS to create two more overall SDTM domains (DM/SUPPDM, EX/SUPPEX) from complete and ongoing oceans, at the same time excluding data from studies identified under point 2 above to reduce their size
4. Add variables in supplemental domains to those already in parent ones and save the resulting datasets as DMPLUS, EXPLUS
5. Call macro %EX\_SUBSET, using DMPLUS, EXPLUS and overall TE, as well as the VC and VF lookup tables, to:
  - a. Identify studies with at least one planned arm containing the vaccine component/formulation of interest
  - b. Ascertain and attach recoded actual exposure information to single subjects (i.e., DMPLUS.ACTARMCD is set to 'A'=exposed/'B'=not exposed)  
NB Subjects not exposed to any study medication (i.e., with no record in EX) are dropped from DMPLUS
6. At this point only data for subjects exposed to the vaccine(s) of interest are kept and double records from extension studies deleted, all in one go:

```
proc sort data = dmplus
    out = _unique_ (where = (actarmcd = 'A'))
    nodupkey;
    by usubjid studyid;
run;
```

```
data _unique_;
    set _unique_;
    by usubjid;
    if first.usubjid;
run;
```

7. The exposure numbers can then be easily calculated and reported by study and sex/race/ age group/country/etc., e.g.:
- ```
proc freq data = _unique_;
    title3"Number of unique subjects in completed or open-label/unblinded
ongoing studies actually exposed at least once to <vaccine>, by study and
sex";
    table studyid * sex /nopercnt nocol norow missing;
run;
```

### PRODUCING A STANDARD LISTING OF SELECTED ADVERSE EVENTS IN DIFFERENT POPULATIONS

This is the high level structure of the second program, and the first five steps are the same as above:

1. Call macro %MRG\_OCEANS to create overall TE domain from complete and ongoing oceans
2. Identify ongoing, blinded studies using the TS domain from ongoing ocean
3. Call again macro %MRG\_OCEANS to create two more overall SDTM domains (DM/SUPPDM, EX/SUPPEX) from complete and ongoing oceans, at the same time excluding data from studies identified under point 2 above to reduce their size
4. Add variables in supplemental domains to those already in parent ones and save the resulting datasets as DMPLUS, EXPLUS

## PhUSE 2015

5. Call macro %EX\_SUBSET, using DMPLUS, EXPLUS and overall TE, as well as the VC and VF lookup tables, to:
  - a. Identify studies with at least one planned arm containing the vaccine component/formulation of interest
  - b. Ascertain and attach recoded actual exposure information to single subjects (i.e., DMPLUS.ACTARMCD is set to 'A'=exposed/'B'=not exposed)  
NB Subjects not exposed to any study medication (i.e., with no record in EX) are dropped from DMPLUS
6. Call macro %MRG\_OCEANS to create a new set of overall SDTM domains (AE, DM, DS, EX, SE plus their supplemental domains, as applicable) from complete and ongoing oceans, this time keeping data only from the exposed subjects identified under point 5.b above
7. Modify accordingly SUPPDM records where QNAM='ACTARM'/'ACTARMD' to either 'A'/'Test' or 'B'/'Control'
8. Create a dataset containing the list of MedDRA terms which will be used in the actual search. Any combination of MedDRA terms from PT to SOC and SMQ (including narrow or broad scope) can be specified, together with the dictionary version to be used (if not the current one)
9. Call macro %GET\_MEDDRA\_TERMS to compile the list of all PTs included in the terms specified above
10. Call macro %SEL\_AE to first subset the AE domain selecting by AEDECOD, then 'sync' SUPPAE. It is consequently possible to use this information to obtain four different scenarios:
  - a. Only selected AEs in subjects exposed the selected vaccine(s)
  - b. All AEs reported in subjects exposed to the selected vaccine(s) and reporting at least one selected AE
  - c. Only selected AEs in all subjects exposed to a study vaccine in studies where at least one vaccination arm contains the selected vaccine(s)
  - d. All AEs in all subjects exposed to a study vaccine in studies where at least one vaccination arm contains the selected vaccine(s)

### MACRO %MRG\_OCEANS

This macro is quite simple, and when called up to two parameters can be specified, the SDTM domain (&DOMAIN) and an optional WHERE statement to subset it (&WHERE\_SUB). The output datasets have the same names as the input ones.

It performs the same actions, first on the parent domain, then, if present, on the supplemental one:

1. If a temporary dataset with that name already exists in WORK, delete it
2. If the dataset exists in the complete ocean, append it to the one in WORK
3. If the dataset exists in the ongoing ocean, append it to the one in WORK
4. Sort resulting dataset by STUDYID and, if not a T-domain, USUBJID, keeping only records with non-missing STUDYID

```
%if %sysfunc(exist(&domain)) %then %do;
proc datasets lib = work
    nolist;
    delete &domain;
run;
%end;
%if %sysfunc(exist(complete.&domain)) %then %do;
proc append base = &domain
    data = complete.&domain (where = (&where_sub));
run;
%end;
%if %sysfunc(exist(ongoing.&domain)) %then %do;
proc append base = &domain
    data = ongoing.&domain (where = (&where_sub))
    force;
run;
%end;

%local add_var;
%if "%substr(&domain, 1, 1)"^="t" %then %let add_var = usubjid;
%else %let add_var = ;

%if %sysfunc(exist(&domain)) %then %do;
proc sort data = &domain (where = (studyid ^= ''))
    force;
    by studyid &add_var;
run;
%end;
```

### MACRO %EX\_SUBSET

This macro accepts up to nine parameters, but two pairs of them are mutually exclusive.

## PhUSE 2015

First all the vaccine formulations of interest are identified by referencing record in either VC or VF, e.g., they all contain a certain adjuvant.

TE is then inspected to compile a list of all the studies with planned use of the selected vaccine(s):

```
PROC SQL NOPRINT;
  CREATE TABLE _study_p AS
  SELECT DISTINCT e.studyid
  FROM   &lib_in..te e, vf_tmp v
  WHERE  INDEXW(COMPRESS(e.element, , 'c'), v.fcode, ' /#') GT 0;
QUIT;
```

Actual exposure is calculated in several ways:

1. all studies where at least one subject has been exposed to the selected vaccine(s)
2. all subjects exposed to any study vaccine in these studies
3. all subjects exposed to the selected vaccine(s)

```
PROC SQL NOPRINT;
  /* Keep all studies with actual exposure to component/vaccine of interest ;
  CREATE TABLE _study_ AS
  SELECT DISTINCT e.studyid
  FROM   &lib_in..explus e, vf_tmp v
  WHERE  INDEXW(COMPRESS(e.extrt, , 'c'), v.fcode, ' /#') GT 0;
  /* Select all subjects exposed to any study vaccine in these studies ;
  CREATE TABLE ex_all AS
  SELECT DISTINCT studyid, usubjid
  FROM   &lib_in..explus
  WHERE  LOWCASE(studyid) IN (&stdy_lst);
  /* Keep subjects with at least one EX record matching any of the VF entries
  selected above, by study ;
  CREATE TABLE ex_tmp AS
  SELECT DISTINCT e.studyid, e.usubjid &add_ex
  FROM   &lib_in..explus e, vf_tmp v
  WHERE  INDEXW(COMPRESS(e.extrt, , 'c'), v.fcode, ' /#') GT 0;
QUIT;
```

As already described earlier, variables DMPLUS.ACTARM/ACTARMCD are then recoded for later use.

### MACRO %GET\_MEDDRA\_TERMS

This macro accepts only one parameter, i.e., the name of the dataset containing the MedDRA version to use and the list of MedDRA terms to be expanded (set by default to WORK.LIST).

This is the template of the short program used to create such datasets:

```
data list (drop = code_);
  infile cards missover pad;
  input TYPE $ 1-4
        CODE_ $ 6-13
        NAME $ 15-134
        ;
  if _n_ = 1 then delete;
  type = upcase(type);
  code = input(code_, 8.);
  name = compbl(upcase(strip(name)));
cards;
**** *****
*****
*****
VERS 17.0
PT 10037778 Radiculitis brachial
PT 10029229 Neuralgic amyotrophy (serious only)
HLT 10043555 Thrombocytopenias
SMQ 20000021 Anaphylactic reaction (narrow)
SMQ 20000024 Angioedema (broad)
;
run;
```

Each term higher than PT (i.e., HLT and above) is expanded into PTs, then the resulting PTs are pooled together and duplicates are deleted.

The macro can set flags in case only serious episodes of a certain MedDRA term are of interest, and use either narrow or broad scope when expanding SMQs.

### MACRO %SEL\_AE

This macro accepts two parameters, controlling if only selected AEs are to be retained (&AESI\_FLG) and if the subjects we are interested into are only those exposed to a certain vaccine or all those enrolled in the selected studies (&SUBSET).

By combining these two options we obtain the four scenarios described earlier.

## PhUSE 2015

```
%if %upcase(&aesi_flg)^=N %then %let where_flg = %str(where = (aesi_flg = 'Y'));
%else %let where_flg = ;

data ae (drop = pt_code pt_ser
         &where_flg);
  merge ae (in = a)
         pt_uniq (rename = (pt_name = aedecod)
                 in = b);
  by aedecod;
  length aesi_flg $ 1;
  aesi_flg = 'N';
  if a;
  /* Check if only serious episodes should be flagged ;
  if b & ^(pt_ser = 'Y' & aeser ^= 'Y') then aesi_flg = 'Y';
run;

%if %upcase(&subset)^=Y %then %do;
%let sub_txt = %str();

/* AEs in selected studies ;
data ae;
  merge ae (in = a)
         _study_ (in = b);
  by studyid;
  if a & b;
run;
%end;
%else %do;
%let sub_txt = %str( in exposed subset);

/* Only AEs in subjects exposed to <vaccine> ;
data ae;
  merge ae (in = a)
         _subset_ (where = (actarmcd = 'A')
                  in = b);
  by studyid usubjid;
  if a & b;
run;
%end;
```

### DATA VISUALIZATION AND CDR

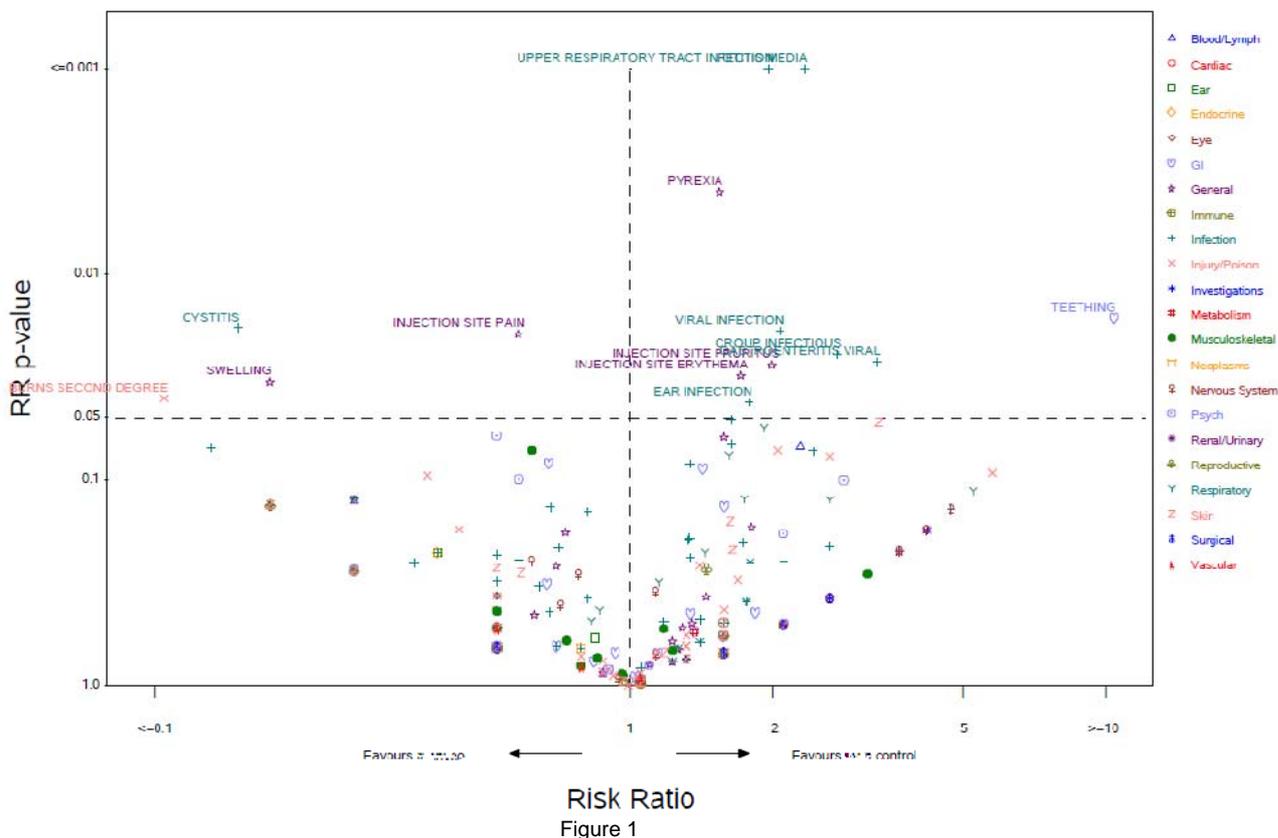
Graphical methods are widely regarded as an excellent way to both explore data for unknown/unexpected correlations as well as analyze large quantities of them together. Given the number of studies now in CDR this is for us a logical step, which we envisioned in our plans from the very beginning anyway.

I had hoped to be able to spend a sizeable amount of time describing our experiences with some commercially available data visualization packages (e.g., JMP Clinical, Spotfire, JReview, etc.) in conjunction with CDR, but technical and contractual issues have made it impossible to do it in time for this paper.

The one aspect linked to graphics that we were able to investigate in my team, even if still to a much lesser extent than initially planned, was that of statistical techniques aimed at safety signal detection. Several SAS programs were developed, aiming at presenting safety data in an aggregated way, e.g., using volcano plots like the one in Figure 1, or forest plots. Research into the use of Bayesian methods to perform safety signaling is also underway, and an internal document was recently finalized, detailing some of the principles we want to follow in these exercises.

Work is still ongoing on these topics, and its developments might be the topic of a further paper in the future.

P-risk (Risk Ratio) Plot of Treatment Emergent Adverse Events at PT Level



**CONCLUSION**

Building on the experience accumulated during the 2009 flu pandemic, it was decided by our management more than 5 years ago that all electronically available data from studies on selected vaccine components or formulations should be concentrated in one physical location using common data structures.

Once that goal was achieved and more than two hundred studies had been pooled it became clear that new opportunities to exploit those data had become suddenly feasible. When I started playing around, showing some early results to some colleagues who were not previously considered as possible clients, they immediately expressed a keen interest. One group of such clients is in Pharmacovigilance, and so we set up a small team to investigate unmet needs.

The two programs I described in detail above, as well as the other ones I only alluded to, are the initial answers to these needs: even if still quite rough the results they produce are highly appreciated, and have shown great promise in reducing the time needed to perform repetitive task, like gathering data for Risk Management Plan (RMP) updates or revising IBs.

**REFERENCES**

Nacci P. Dynamic Oceans and Static Poolings of Clinical Trial Data in NVx. Proceedings of the PhUSE 2014 Annual Conference, London

**ACKNOWLEDGMENTS**

I would like to thank Francesca Dall'Acqua, Elizabeth Merrall and Nicholas Roubinis for their precious comments and support for the data visualization section.

**CONTACT INFORMATION**

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

Pantaleo Nacci  
 GSK Vaccines S.r.l.  
 Via Fiorentina, 1  
 53100 Siena, Italy  
 Phone: +39 0577 243554  
 Email: [pantaleo.x.nacci@gsk.com](mailto:pantaleo.x.nacci@gsk.com)  
 Web: <http://www.gsk.com/>

Brand and product names are trademarks of their respective companies.

