PhUSE 2014

Paper DH10

Dynamic Oceans and Static Poolings of Clinical Trial Data in NVx

Pantaleo Nacci, Novartis Vaccines & Diagnostics Srl, Siena, Italy

ABSTRACT

Business drivers to merge legacy and current clinical trials data using common, non-proprietary data structures are increasing, often giving a competitive advantage to companies which can query all data available from clinical trials. Novartis Vaccines has based the decision to develop its Clinical Data Repository on this aspect, as well as other important ones. This paper gives some insights on how data from multiple trials are assembled and utilized to support regulatory submissions, PSUR/DSUR/IB preparation, etc. Special attention is paid to the mechanism devised to allow the identification of what has been administered to each subject.

INTRODUCTION

As part of the much wider Clinical Data Repository project, at Novartis Vaccines we have been quite busy for the last three years remapping data from available legacy clinical trials (as well as of course collecting data from new ones) according to a set of common, non-proprietary data standards from CDISC. Initially we adopted just CDASH and SDTM, but we have recently moved into ADaM as well.

Now that SDTM data for more than 280 clinical trials are available, the challenge is to make sure that they are easily and reliably accessible, so as to support various types of analyses including data from multiple studies. Last but not least, study metadata need to be correctly managed to avoid studies being lost among the throng.

CDR - A QUICK RECAP

Clinical Data Repository (CDR) is the name of the Novartis Vaccines (NVx) 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

BRINGING ALL TOGETHER

Figure 1 on the next page is the graphical representation of how the CDR is placed in the context of our revised data flows, playing the role of central hub for many processes previously linked much more loosely.

E.g., in previous EDC studies the coding of adverse events, medical history, medications, etc., was managed on a study-by-study basis, thus creating huge issues when there was the need to unify dictionary-level terminology after pooling of data. That task has now been unified so that all coding is performed centrally in Oracle®'s Thesaurus Management System (TMS).

There are still further developments waiting in the wings, like a tighter integration of our Pharmacovigilance systems and a graphical data exploration tool. On top of that other functions have expressed interest in getting (mostly read-only) access to CDR data for their own activities.

BUSINESS DRIVERS

The main business drivers for its implementation were:

- Increase in volume and complexity of Health Authorities expectations (e.g., 2009 flu pandemic)
- Reduced turnaround times for questions, often involving multiple studies/projects
- Number of studies and subjects going up, increasing the workload
- Push to reduce time needed to develop a new vaccine while maintaining quality
- Ultimately, need to get rid of outdated, non-scalable processes & systems

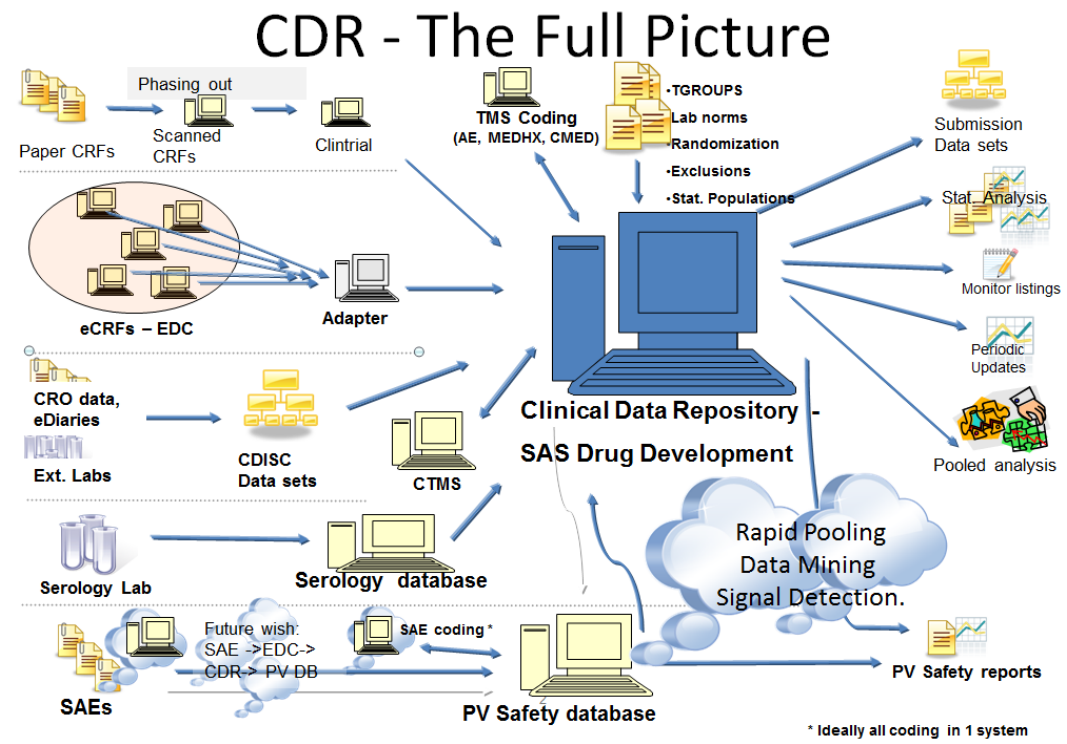


Figure 1: The CDR as a hub

STUDY DIRECTORIES

A single file (`all_trials.xls`) is maintained centrally (Fig. 2), containing details about all studies known to CDR, including their origin (legacy or CDR) and execution status (active or completed). Each study is allocated its own set of directories to store data, differentiated accordingly to this information.

	A	B	C	D	E	H
1	Obs	CDPNAME	STUDYID	ACTIVE	SOURCE	AREA
874	864	V59	V59P13E1	NO	CLINTRIAL	PROD
875	865	V59	V59P14E1	NO	EDC-CIS	PROD
876	866	V102	V102_03	NO	EDC-CIS	PROD
877	867	V102	V102_02E1	NO	CLINTRIAL	PROD
881	871	V72	V72_60	YES	EDC	PROD
883	873	V102	V102_15	YES	EDC	PROD

Figure 2: The all_trials.xls file

For legacy studies, directory `/data/<cdpname>/<studyid>/operational/prod/ssd` contains the original, unmapped data (Fig. 3), while the corresponding SDTM version is stored under `/data/<cdpname>/<studyid>/ snapshots/prod/pool/sdtm` (Fig. 4).

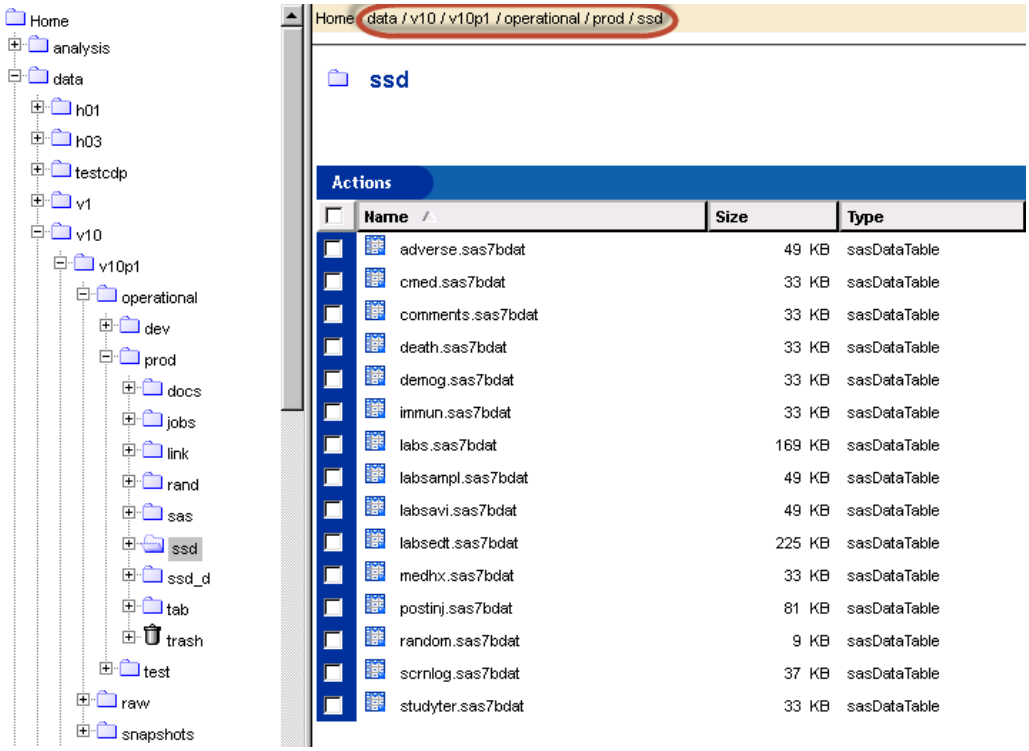


Figure 3: Legacy trial, original data

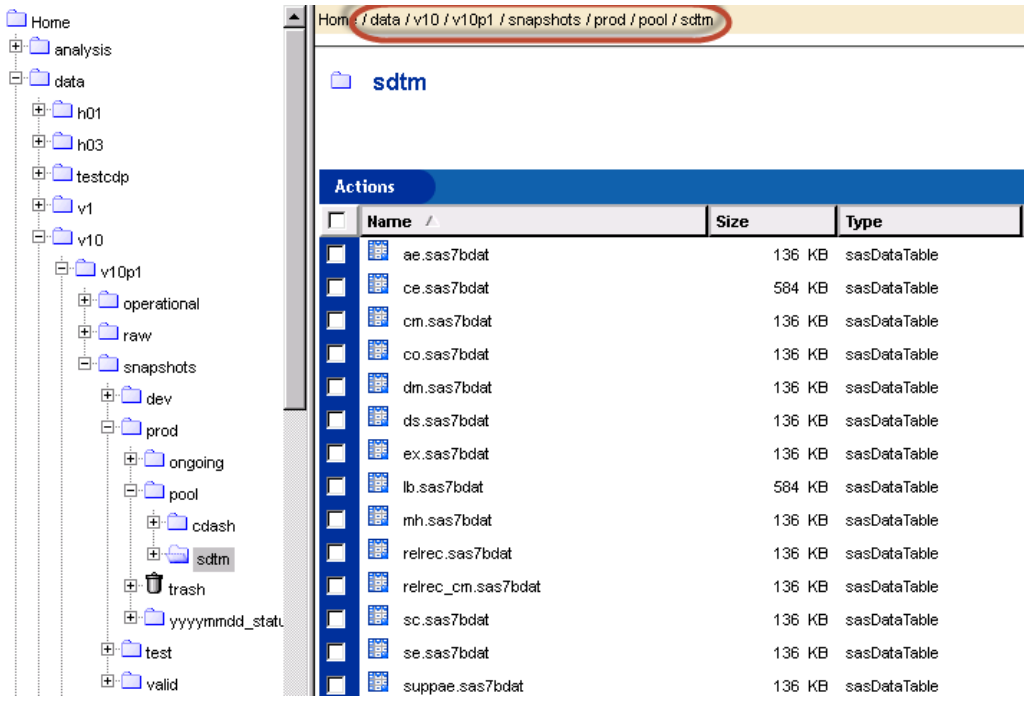


Figure 4: Legacy trial, SDTM remapped data

For CDR studies, directory `/data/<cdpname>/<studyid>/snapshots/prod/<status>/cdash` contains the CDASH data as collected in the EDC system (Fig. 5), while the corresponding SDTM version is stored under `/data/<cdpname>/<studyid>/snapshots/prod/<status>/sdtm` (Fig. 6). The value of `<status>` can be either `ongoing` (active studies) or `pool` (completed studies).

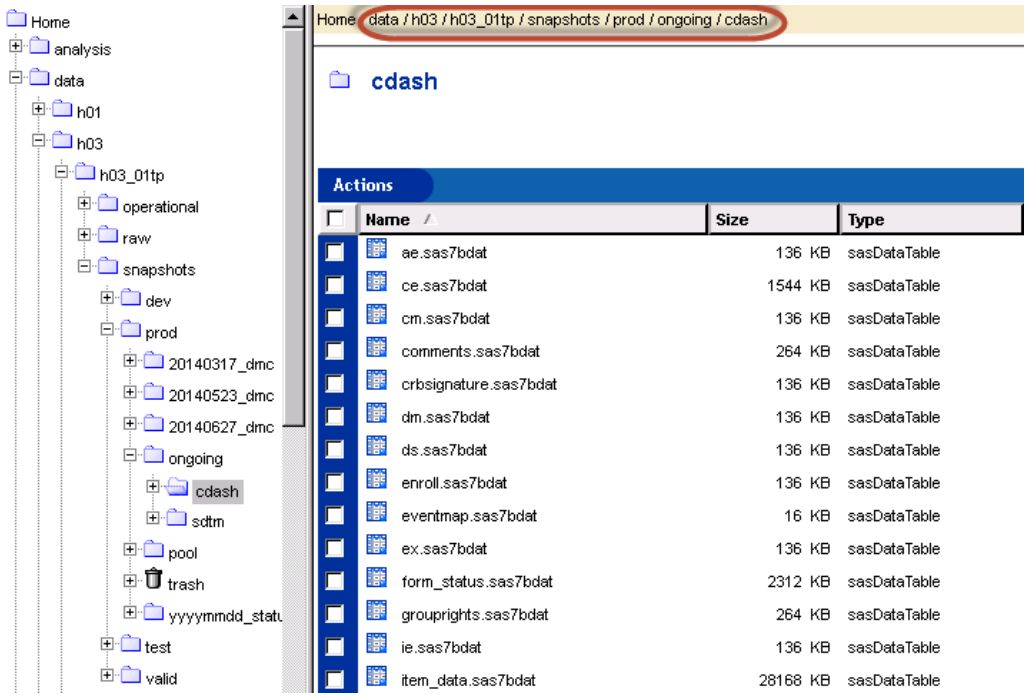


Figure 5: CDR ongoing trial, CDASH data

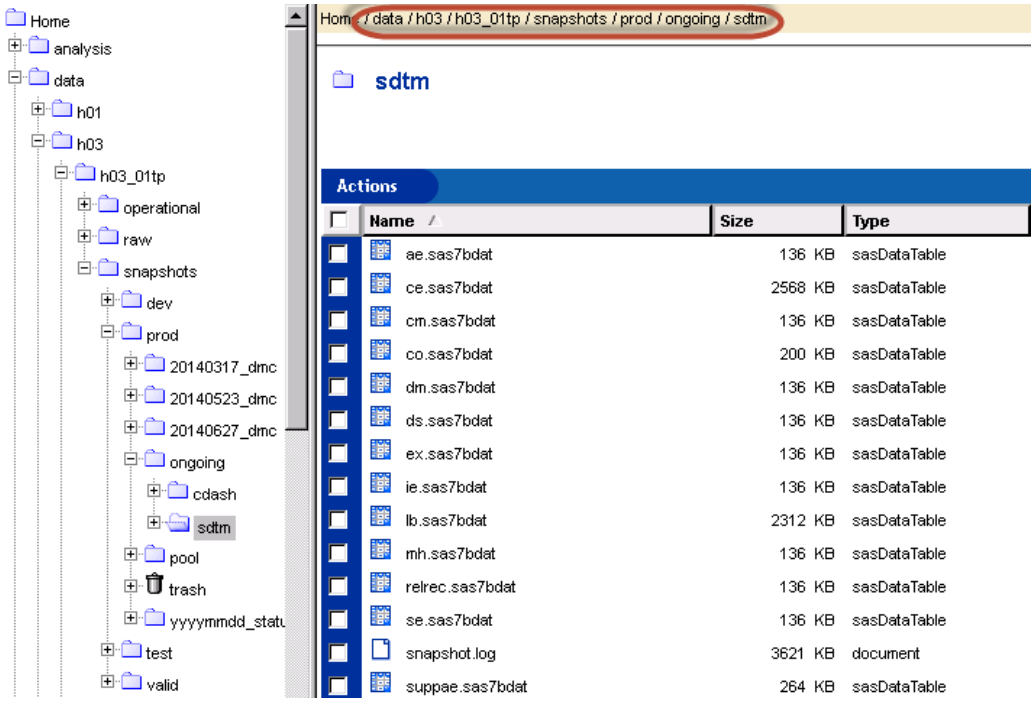


Figure 6: CDR ongoing trial, SDTM data

This standard directory structure gives the opportunity to write programs which are completely driven by the data stored in `all_trials.xls` above.

Data in these directories contain the coding for adverse events, medical history and non-study medication performed at the time the data were originally extracted, so, e.g., a legacy study would have its AEs coded using an early version of MedDRA®, or even in COSTART.

(DYNAMIC) DATA OCEANS

The inhomogeneity is of course a major issue whenever analyses spanning multiple trials are needed, so we came up with the concept of data 'oceans', to collect in one place (actually two) the same data stored in the single study directories, but with selected variables (e.g., `--DECOD`, `--BODSYS`) recoded according to the same dictionary, including its version.

When data for a study are available in CDR, they are added to either one of two oceans (`/analysis/pool/<ocean>/prod/sdtm`, Fig. 7), according to its status in `all_trials.xls`:

- Data from active/ongoing studies are added to the `ongoing` ocean

- Data from a legacy or complete study are added to the `complete` ocean

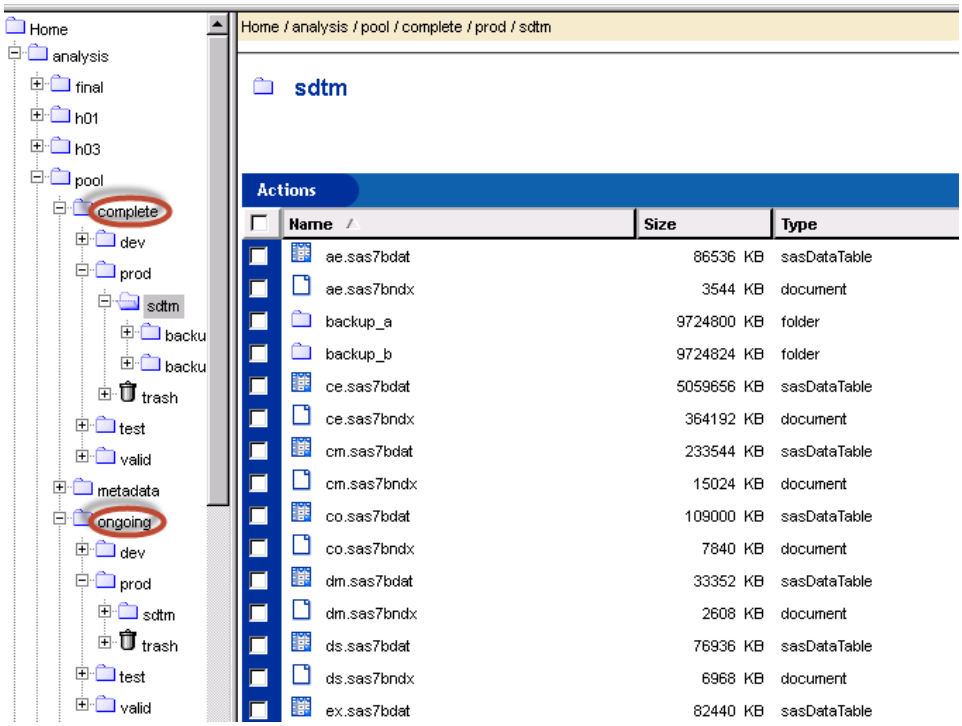


Figure 7: The data oceans

The logic used to create and maintain the two oceans is very different: while data from completed studies are added over time to the complete ocean in an incremental fashion, the ongoing one is recreated from scratch every night. In case of post DB lock changes there is a documented process to be followed to refresh the interested data in the complete ocean. At any given time data for a study can be present in only one ocean. Last but not least it is possible to update all MedDRA terminology at once whenever a new version is implemented. These three tasks (i.e., maintaining the complete ocean, creating the ongoing one, and updating MedDRA terminology) are managed by one validated SAS utility, which can be executed in three corresponding modes (Fig. 8). The utility is currently scheduled to run in both `ongoing` and `complete` mode every night, while it is manually run in `meddra` mode twice every year.

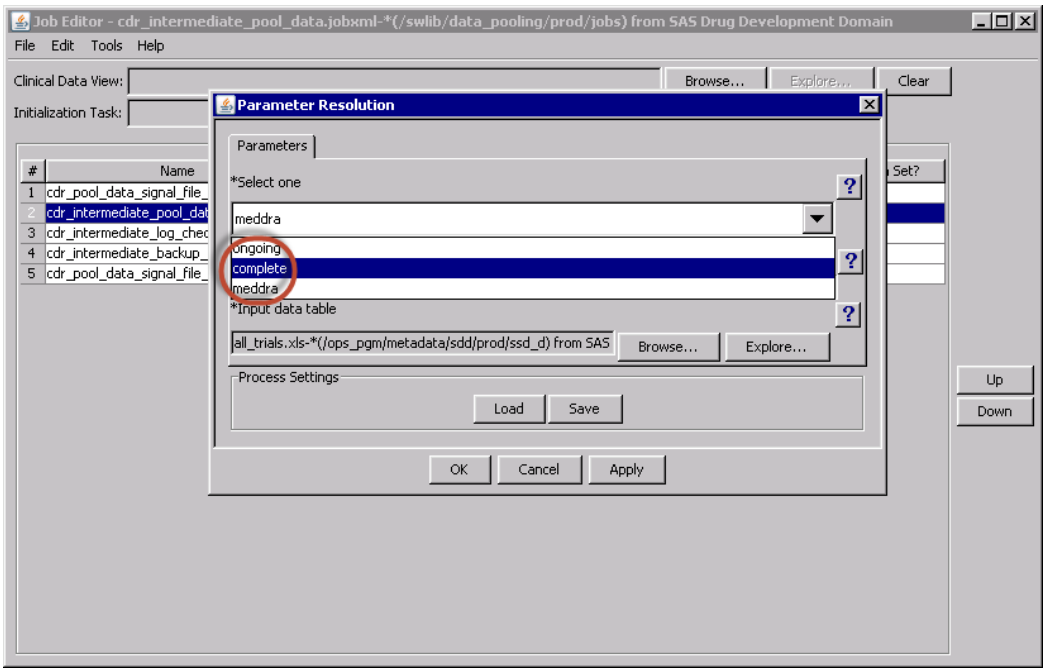
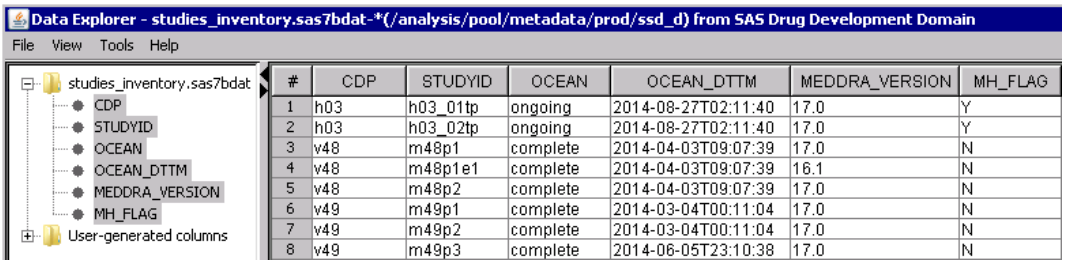


Figure 8: The data ocean SAS utility

Information about which studies are included in each ocean is stored in the metadata file `studies_inventory` (Fig. 9), while a full list of which domains are present for each study in an ocean are stored in `ongoing_inventory` and `complete_inventory` (Fig. 10).

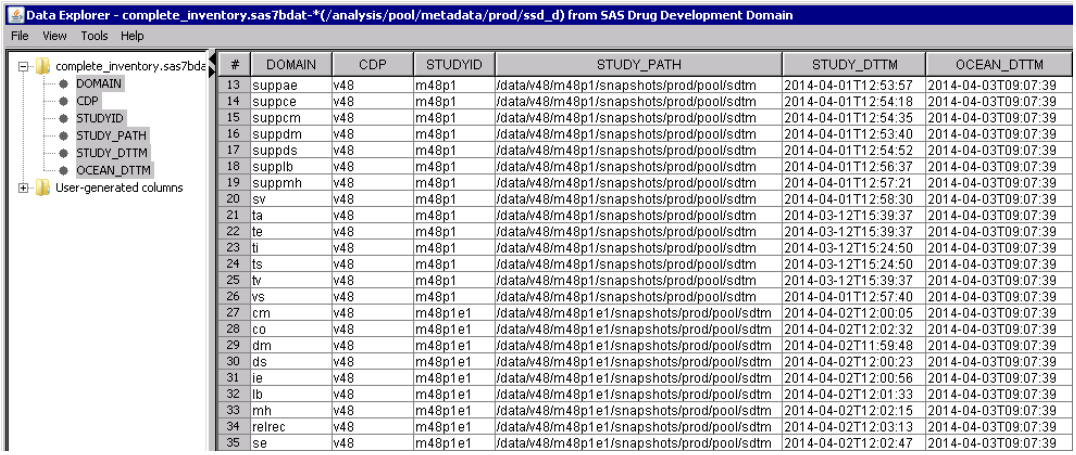

Data Explorer - studies_inventory.sas7bdat-*(/analysis/pool/metadata/prod/ssd_d) from SAS Drug Development Domain

File View Tools Help

studies_inventory.sas7bdat
CDP
STUDYID
OCEAN
OCEAN_DTTM
MEDDRA_VERSION
MH_FLAG
User-generated columns

#	CDP	STUDYID	OCEAN	OCEAN_DTTM	MEDDRA_VERSION	MH_FLAG
1	h03	h03_01tp	ongoing	2014-08-27T02:11:40	17.0	Y
2	h03	h03_02tp	ongoing	2014-08-27T02:11:40	17.0	Y
3	v48	m48p1	complete	2014-04-03T09:07:39	17.0	N
4	v48	m48p1e1	complete	2014-04-03T09:07:39	16.1	N
5	v48	m48p2	complete	2014-04-03T09:07:39	17.0	N
6	v49	m49p1	complete	2014-03-04T00:11:04	17.0	N
7	v49	m49p2	complete	2014-03-04T00:11:04	17.0	N
8	v49	m49p3	complete	2014-06-05T23:10:38	17.0	N

Figure 9: The metadata dataset studies_inventory


Data Explorer - complete_inventory.sas7bdat-*(/analysis/pool/metadata/prod/ssd_d) from SAS Drug Development Domain

File View Tools Help

complete_inventory.sas7bda
DOMAIN
CDP
STUDYID
STUDY_PATH
STUDY_DTTM
OCEAN_DTTM
User-generated columns

#	DOMAIN	CDP	STUDYID	STUDY_PATH	STUDY_DTTM	OCEAN_DTTM
13	suppae	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-04-01T12:53:57	2014-04-03T09:07:39
14	suppce	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-04-01T12:54:18	2014-04-03T09:07:39
15	suppcm	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-04-01T12:54:35	2014-04-03T09:07:39
16	suppdm	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-04-01T12:53:40	2014-04-03T09:07:39
17	suppds	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-04-01T12:54:52	2014-04-03T09:07:39
18	supplb	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-04-01T12:56:37	2014-04-03T09:07:39
19	suppmh	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-04-01T12:57:21	2014-04-03T09:07:39
20	sv	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-04-01T12:58:30	2014-04-03T09:07:39
21	ta	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-03-12T15:39:37	2014-04-03T09:07:39
22	te	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-03-12T15:39:37	2014-04-03T09:07:39
23	ti	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-03-12T15:24:50	2014-04-03T09:07:39
24	ts	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-03-12T15:24:50	2014-04-03T09:07:39
25	tv	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-03-12T15:39:37	2014-04-03T09:07:39
26	vs	v48	m48p1	/data/v48/m48p1/snapshots/prod/pool/sdtm	2014-04-01T12:57:40	2014-04-03T09:07:39
27	cm	v48	m48p1e1	/data/v48/m48p1e1/snapshots/prod/pool/sdtm	2014-04-02T12:00:05	2014-04-03T09:07:39
28	co	v48	m48p1e1	/data/v48/m48p1e1/snapshots/prod/pool/sdtm	2014-04-02T12:02:32	2014-04-03T09:07:39
29	dm	v48	m48p1e1	/data/v48/m48p1e1/snapshots/prod/pool/sdtm	2014-04-02T11:59:48	2014-04-03T09:07:39
30	ds	v48	m48p1e1	/data/v48/m48p1e1/snapshots/prod/pool/sdtm	2014-04-02T12:00:23	2014-04-03T09:07:39
31	ie	v48	m48p1e1	/data/v48/m48p1e1/snapshots/prod/pool/sdtm	2014-04-02T12:00:56	2014-04-03T09:07:39
32	lb	v48	m48p1e1	/data/v48/m48p1e1/snapshots/prod/pool/sdtm	2014-04-02T12:01:33	2014-04-03T09:07:39
33	mh	v48	m48p1e1	/data/v48/m48p1e1/snapshots/prod/pool/sdtm	2014-04-02T12:02:15	2014-04-03T09:07:39
34	relrec	v48	m48p1e1	/data/v48/m48p1e1/snapshots/prod/pool/sdtm	2014-04-02T12:03:13	2014-04-03T09:07:39
35	se	v48	m48p1e1	/data/v48/m48p1e1/snapshots/prod/pool/sdtm	2014-04-02T12:02:47	2014-04-03T09:07:39

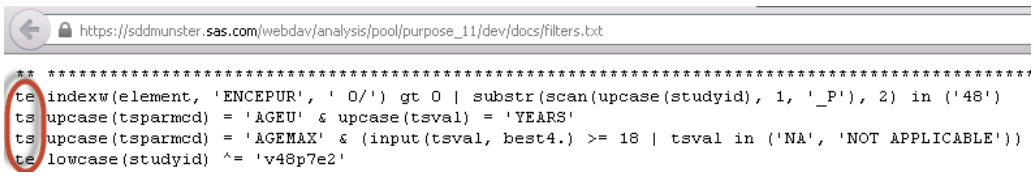
Figure 10: The metadata dataset complete_inventory

(STATIC) DATA POOLINGS

Since they are modified every night, at least theoretically, the oceans are not optimal to support analyses where the ability to replicate existing outputs or create new ones months after the fact is critical.

To accommodate such needs another validated SAS utility was written, which can create static snapshots of the combined oceans on a certain date, usually containing only a subset of available data, which we call data poolings.

The content of a data pooling is established by means of a simple ASCII file (`filters.txt`), containing one or more SAS expressions involving only T-domains and linked in a logical 'AND' fashion, i.e., all conditions must be met for a study to be included (Fig. 11).


https://sddmunster.sas.com/webdav/analysis/pool/purpose_11/dev/docs/filters.txt

```
** ************************************************************************************************;
te indexw(element, 'ENCEPUR', ' O/') gt 0 | substr(scan(upcase(studyid), 1, '_P'), 2) in ('48')
ts upcase(tsparmcd) = 'AGEU' & upcase(tsval) = 'YEARS'
ts upcase(tsparmcd) = 'AGEMAX' & (input(tsval, best4.) >= 18 | tsval in ('NA', 'NOT APPLICABLE'))
te lowcase(studyid) ^= 'v48p7e2'
```

Figure 11: An example of `filters.txt` file

The utility can be run in two modes, the first (`Filter Check`) to test that the right studies are selected and adjust the contents of `filters.txt` accordingly, the other (`Create Pool`) to physically create the data pooling itself (Fig. 12).

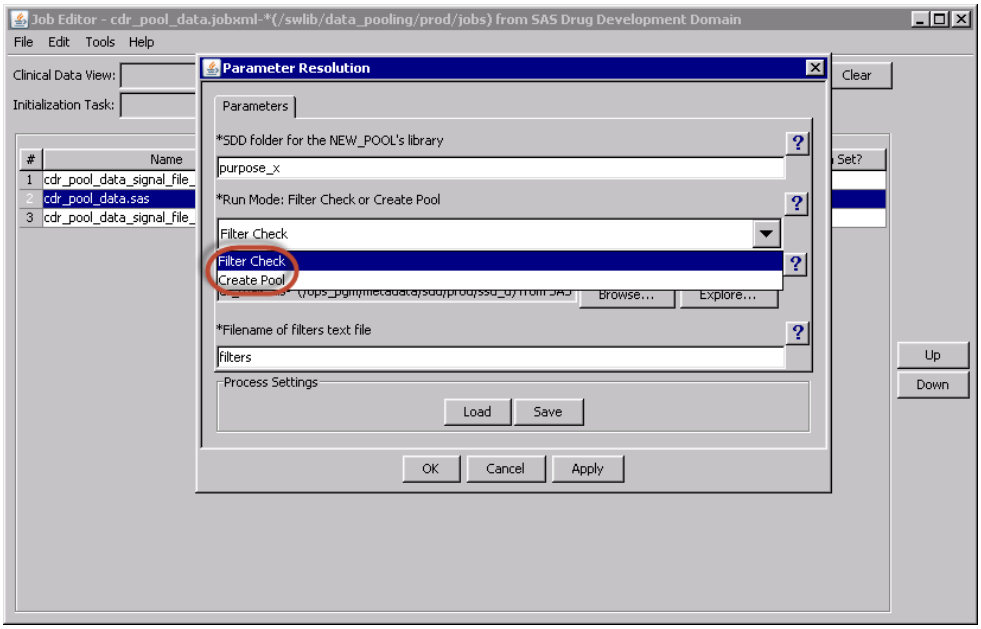


Figure 12: The data pooling SAS utility

Each data pooling is stored under its own directory `/analysis/pool/purpose_<x>/prod/sdtm` (Fig. 13), and contains all data available for the studies selected in the `filters.txt` file. If only a part of the data is actually needed, this further subject-level subsetting will be performed at analysis time.

As the number of available studies increases, it is more and more difficult to predict in advance which studies should be selected, e.g., when a vaccine has been used as active placebo in another project. It is thus extremely important that metadata are double checked for each study, to avoid losing a study along the way due to wrong information in its T-domains.

Last but not least, once a data pooling has been created in production the utility is by design unable to overwrite it.

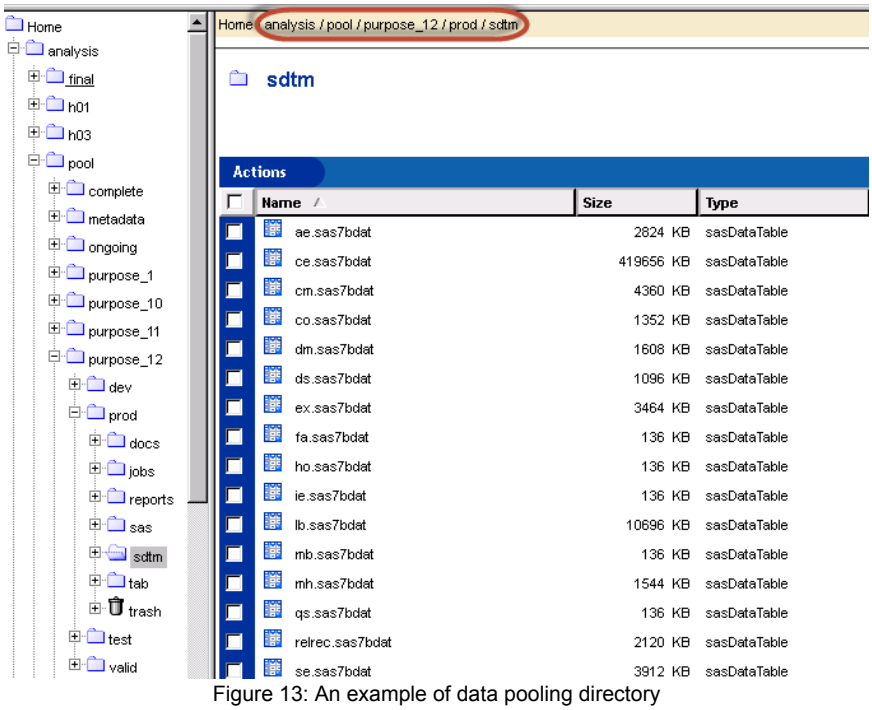


Figure 13: An example of data pooling directory

HOW EXPOSURE INFORMATION IS STORED IN CDR

One of the most interesting characteristics we built into CDR is the ability to easily identify very specific populations, e.g., all subjects less than 6 month of age at enrolment exposed to at least one dose of any thiomersal-containing flu vaccine in Canada.

To help in running such queries we designed a couple of custom metadata domains or lookup tables, to identify precisely what lies behind a TA.ELEMENT (or EX.EXTRT) value (Fig. 14).

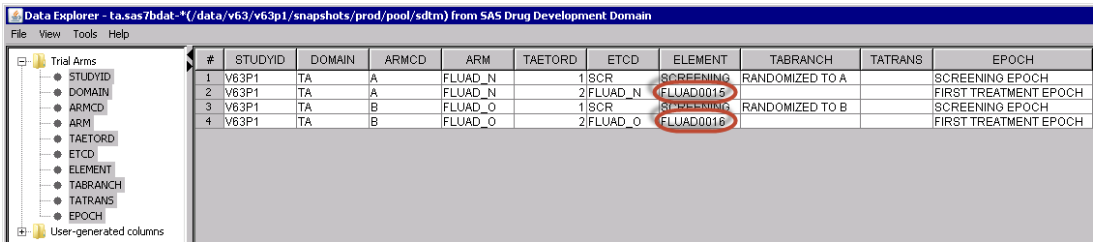


Figure 14: An example of TA domain

The first one, VC ('Vaccine Components'), contains the basic 'building blocks' of all the vaccines we have used in our clinical trials until now (Figure 15), while the other, VF ('Vaccine Formulations'), uses a subset of VC records to represent the exact composition of a certain vaccine (Figure 16).

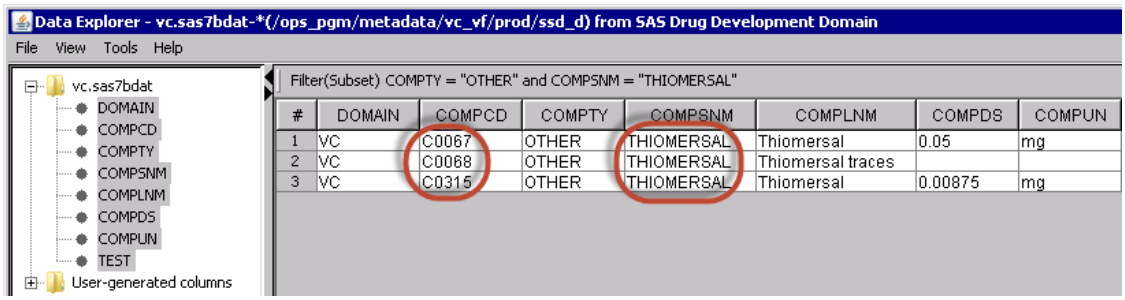


Figure 15: The VC lookup table

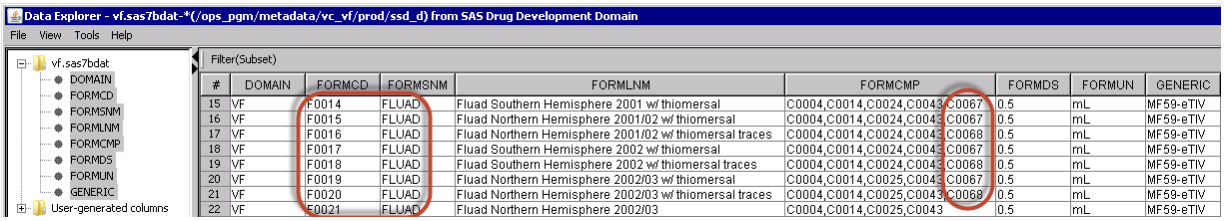


Figure 16: The VF lookup table

This logical structure allows for a high degree of flexibility, since VC elements can be reused in unlimited permutations. Searches can be written looking at VF only (e.g., all formulations of a certain vaccine) or at VC (e.g., all vaccines containing a certain antigen).

DIFFERENT ANALYSIS STRATEGIES USING OCEANS AND DATA POOLINGS

As briefly mentioned above already, different analysis strategies might need to use either oceans or data poolings.

ANALYSES USING OCEANS

Oceans have been used as data source for analyses which are run routinely over time, but where there is no need to replicate earlier results, since they have a meaning only at a very specific point in time.

Examples of such analyses are the periodic listings of clinical trials adverse events which we create for our Pharmacovigilance colleagues or the numbers of subject exposed to a certain vaccine as of a certain date.

Usually all available data at the time the analysis is run are used, with the logical exceptions (e.g., data from ongoing blinded trials cannot usually be counted on to calculate exposure numbers).

ANALYSES USING DATA POOLINGS

On the other hand data poolings are created when we need to be able to replicate the same results at a later time, or create further variations of them.

Thus they include mostly regulatory submissions (BLAs, MMAs), DSUPs, PSURs, IBs, etc., but also journal papers.

The subset of studies to be included is in general already defined in the Statistical Analysis Plan, so that data from ongoing studies are there only when planned.

CONCLUSION

The number of studies available in the same data format in CDR is steadily increasing, and we have only started to glimpse at what new analysis strategies we have now available at our fingertips.

We have strived to avoid duplicating information in multiple places, so we have devised a logical structure allowing us to write as much as possible purely data-driven programs, not needing continuous modifications. To this end we spent a lot of time and effort in coming up with a directory structure at the same time highly standardised and flexible.

The oceans allow us to have all data in one physical local and to recode selected dictionary-related variable to make data completely homogeneous, while data poolings give us the opportunity to save static snapshots of just the studies we need to support a certain analysis.
The importance and relevance of correct study metadata is also being highlighted over and over, so a shift in approach is needed there to be able to fully exploit the benefits of a system like CDR.

CONTACT INFORMATION

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

Pantaleo Nacci
Novartis Vaccines & Diagnostics Srl
Via Fiorentina, 1
53100 Siena, Italy
Phone: +39 0577 243554
Fax: +39 0577 278443
Email: pantaleo.nacci@novartis.com
Web: www.novartisvaccines.com