



Dynamic Oceans and Static Poolings of Clinical Trial Data in NVx

Pantaleo Nacci, Head Statistical Safety & Epidemiology/PV

PhUSE Annual Conference

London, 13 October 2014



Agenda



CDR- A Quick Recap

One File to List Them All

(Dynamic) Data Oceans

(Static) Data Poolings

How Exposure Information is Stored

Conclusions



CDR – A Quick Recap

A move towards data standardisation and integration

- 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)
- 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

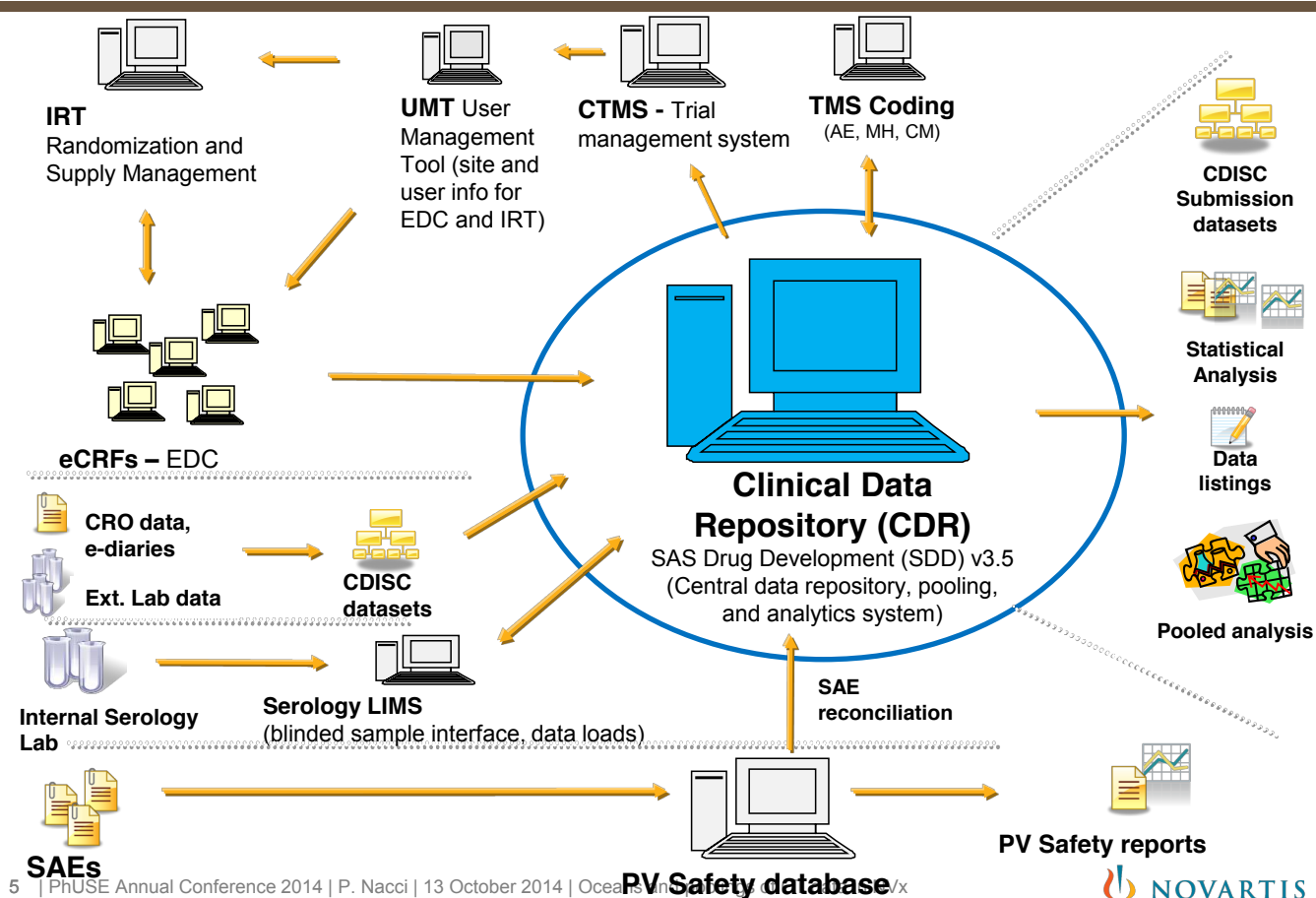
Why a Clinical Data Repository?

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 - Ultimately, need to get rid of outdated, non-scalable processes & systems

CDR Overview

A hub for multiple processes



Agenda



CDR – A Quick Recap

One File to List Them All

(Dynamic) Data Oceans

(Static) Data Poolings

How Exposure Information is Stored

Conclusions



One File to List Them All

Avoiding data replication

- CDR is designed to manage data (and metadata) from hundreds of studies
 - A coherent, logical system of directories was needed
- Replication of metadata (as well as data, to a lesser extent) is avoided as much as possible
- The need for a 'master file' of all studies in the system was identified early on
 - If a study is not present in there, for all intents and purposes it does not exist
- The solution is a single Excel file (`all_trials.xls`), containing information about all studies

Excerpt from `all_trials.xls`

Unique source of study data information

- The study info includes origin (legacy or CDR-native), execution status (active or completed), and more
- 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

Directory Structure - Study Data

Standardised as to be highly programmable

■ For legacy studies:

- /data/<cdpname>/<studyid>/operational/prod/ssd contains the original, unmapped data
- /data/<cdpname>/<studyid>/snapshots/prod/pool/sdtm stores the corresponding SDTM version

■ For CDR-native studies:

- /data/<cdpname>/<studyid>/snapshots/prod/<status>/cdash contains the CDASH data as collected in the EDC system
- /data/<cdpname>/<studyid>/snapshots/prod/<status>/sdtm stores the corresponding SDTM version
 - The value of <status> can be either ongoing (active studies) or pool (completed studies)

Dictionary Coding is Still the Original One

Replicating original analyses is possible

- Final SDTM data for all completed studies are always found under the same directory, no matter the origin
- This structure gives the opportunity to write programs which are completely driven by the data stored in `all_trials.xls` above
- Data for adverse events (AE), medical history (MH) and non-study medication (CM) are coded as at the time they were originally finalised/extracted
 - E.g., a legacy study from 2001 would have its AEs coded using an earlier version of MedDRA, if not in COSTART

Agenda



CDR – A Quick Recap

One File to List Them All

(Dynamic) Data Oceans

(Static) Data Poolings

How Exposure Information is Stored

Conclusions



(Dynamic) Data Oceans

Where all available study data can be found

- The inhomogeneity is a major issue whenever analyses spanning multiple trials are planned
- We then came up with the concept of data ‘oceans’
 - Collect in one place the data stored in the single study directories
 - Selected AE, CM and MH variables (e.g., --DECOD, --BODSYS) are recoded to make them homogeneous
- When data are available in CDR for a study, they are added to either one of two oceans, according to its status in `all_trials.xls`:
 - `/analysis/pool/<ocean>/prod/sdtm`
 - The value of `<ocean>` can be either `ongoing` (active studies) or `complete` (completed studies)

Logic Used to Maintain Data Oceans

Different according to study status

- The logic used to create and maintain the two oceans is different:
 - data from newly completed studies are added 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
 - 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 in the oceans in one go whenever a new version is implemented

A Single Utility to Perform all Tasks

Changing a program parameter is all that is needed

- All tasks are managed by one validated SAS utility
- The utility is currently scheduled to run
 - automatically in both `ongoing` and `complete` mode every night
 - manually in `meddra` mode twice every year
- Information about which studies are included in each ocean is stored in the metadata file `studies_inventory`
- A list of which domains are present for each study in an ocean are stored in `ongoing_inventory` and `complete_inventory`

Analysis Scenarios – Data Oceans

Actual examples of data oceans use

- Analyses using oceans as data sources are run repeatedly over time
 - No need to replicate earlier results
 - Results meaningful only at a very specific point in time
- Examples:
 - periodic listings of specific adverse events for Pharmacovigilance
 - 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 logical exceptions
 - E.g., data from ongoing blinded trials cannot usually be counted on to calculate exposure numbers

Agenda



CDR – A Quick Recap

One File to List Them All

(Dynamic) Data Oceans

(Static) Data Poolings

How Exposure Information is Stored

Conclusions



(Static) Data Poolings

Subset of study data needed for a specific purpose

- Data oceans are modified every night → Not optimal to support analyses where the ability to replicate existing outputs or create new variations of them is critical
- Another SAS utility was developed and validated
 - Create static snapshots ('data poolings') of the combined oceans on a certain date
 - They contain all available data for a subset of studies
- Contents of a data pooling are defined in an ASCII file (`filters.txt`), listing one or more SAS expressions
 - Only variables in T-domains are usable in expressions
 - All expressions are linked in a logical 'AND' fashion
 - All conditions must be met for a study to be included

Another SAS Utility

Testing the contents before creating a data pooling

- This utility can be run in two modes:
 - The first (`Filter Check`) tests that right studies are selected
 - Contents of `filters.txt` are adjusted as needed
 - The second (`Create Pool`) creates the data pooling itself
- Each data pooling is stored under its own directory:
 - `/analysis/pool/purpose_<x>/prod/sdtm`
 - `<x>` is a number, increasing every time a new data pooling is requested
- Once a data pooling has been created in production the utility is unable to overwrite it by design

Making Sure All Right Studies Are Selected

It is possible to loose a study in the oceans

- As the number of studies available in CDR increases, it is more and more difficult to know in advance which studies should be selected
 - E.g., vaccines used as active placebos in other projects
- It is thus extremely important that metadata are double checked for each study
 - There is a real risk of 'losing' a study due to wrong information
- Subject-level subsetting is performed at analysis time
 - E.g., if we are only interested in paediatric data from a study recruiting adolescents too

Analysis Scenarios – Data Poolings

Actual examples of data poolings use

- Data poolings are created when it is necessary to be able to replicate the same results at a later time, or to create further variations of them
- Thus they mostly include
 - Regulatory submissions (BLAs, MMAs)
 - DSUPs, PSURs, IBs, etc.
 - But also journal papers
- The subset of studies to be included is defined in the Statistical Analysis Plan
 - Data from ongoing studies are included only when specifically mentioned

Agenda



CDR – A Quick Recap

One File to List Them All

(Dynamic) Data Oceans

(Static) Data Poolings

How Exposure Information is Stored

Conclusions



How Exposure Information is Stored

Understanding exactly who was exposed to what

- We built into CDR the ability to identify specific populations
 - E.g., all Canadian subjects less than 18 years of age at enrolment exposed to at least one dose of any thiomersal-containing vaccine
- To deal with substance exposure we designed a couple of custom metadata domains/lookup tables
 - E.g., identify what is behind a TA.ELEMENT (or EX.EXTRT) value

Data Explorer - ta.sas7bdat-*(//data/v63/v63p1/snapshots/prod/pool/sdtm) from SAS Drug Development Domain

File View Tools Help

#	STUDYID	DOMAIN	ARMCD	ARM	TAETORD	ETCD	ELEMENT	TABRANCH	TATRANS	EPOCH
1	V63P1	TA	A	FLUAD_N	1	SCR	SCREENING	RANDOMIZED TO A		SCREENING EPOCH
2	V63P1	TA	A	FLUAD_N	2	FLUAD_N	FLUAD0015			FIRST TREATMENT EPOCH
3	V63P1	TA	B	FLUAD_O	1	SCR	SCREENING	RANDOMIZED TO B		SCREENING EPOCH
4	V63P1	TA	B	FLUAD_O	2	FLUAD_O	FLUAD0016			FIRST TREATMENT EPOCH

Left sidebar: Trial Arms

- STUDYID
- DOMAIN
- ARMCD
- ARM
- TAETORD
- ETCD
- ELEMENT
- TABRANCH
- TATRANS
- EPOCH

User-generated columns

Two-level Approach

Specifying single components within each vaccine formulation

- The first one, VC ('Vaccine Components'), contains the 'building blocks' of vaccines we used in our clinical trials
 - This includes antigens, adjuvants, additives, etc.
 - Elements can be reused
- The other, VF ('Vaccine Formulations'), uses a subset of VC records to represent the exact composition of a certain vaccine, i.e., its formulation
- This two-tier structure allows for a high degree of flexibility
- Searches can then be written looking at VF (e.g., all formulations of a certain vaccine) or at VC (e.g., all vaccines containing a certain component)

What is Behind an ELEMENT

Example: 'FLUAD0015', 'FLUAD0016'

Data Explorer - vf.sas7bdat-*(/ops_pgm/metadata/vc_vf/prod/ssd_d) from SAS Drug Development Domain

File View Tools Help

Filter(Subset)

#	DOMAIN	FORMCD	FORMSNM	FORMLNM	FORMCMP	FORMDS	FORMUN	GENERIC
15	VF	F0014	FLUAD	Fluad Southern Hemisphere 2001 w/ thiomersal	C0004,C0014,C0024,C0043,C0067	0.5	mL	MF59-eTIV
16	VF	F0015	FLUAD	Fluad Northern Hemisphere 2001/02 w/ thiomersal	C0004,C0014,C0024,C0043,C0067	0.5	mL	MF59-eTIV
17	VF	F0016	FLUAD	Fluad Northern Hemisphere 2001/02 w/ thiomersal traces	C0004,C0014,C0024,C0043,C0068	0.5	mL	MF59-eTIV
18	VF	F0017	FLUAD	Fluad Southern Hemisphere 2002 w/ thiomersal	C0004,C0014,C0024,C0043,C0067	0.5	mL	MF59-eTIV
19	VF	F0018	FLUAD	Fluad Southern Hemisphere 2002 w/ thiomersal traces	C0004,C0014,C0024,C0043,C0068	0.5	mL	MF59-eTIV
20	VF	F0019	FLUAD	Fluad Northern Hemisphere 2002/03 w/ thiomersal	C0004,C0014,C0025,C0043,C0067	0.5	mL	MF59-eTIV
21	VF	F0020	FLUAD	Fluad Northern Hemisphere 2002/03 w/ thiomersal traces	C0004,C0014,C0025,C0043,C0068	0.5	mL	MF59-eTIV
22	VF	F0021	FLUAD	Fluad Northern Hemisphere 2002/03	C0004,C0014,C0025,C0043	0.5	mL	MF59-eTIV

User-generated columns

Data Explorer - vc.sas7bdat-*(/ops_pgm/metadata/vc_vf/prod/ssd_d) from SAS Drug Development Domain

File View Tools Help

Filter(Subset) COMPTY = "OTHER" and COMPSNM = "THIOMERSAL"

#	DOMAIN	COMPCD	COMPTY	COMPSNM	COMPLNM	COMPDS	COMPUN
1	VC	C0067	OTHER	THIOMERSAL	Thiomersal	0.05	mg
2	VC	C0068	OTHER	THIOMERSAL	Thiomersal traces		
3	VC	C0315	OTHER	THIOMERSAL	Thiomersal	0.00875	mg

User-generated columns

Agenda



CDR – A Quick Recap

One File to List Them All

(Dynamic) Data Oceans

(Static) Data Poolings

How Exposure Information is Stored

Conclusions



Conclusions

Summing up

- The number of studies available in the same data format in CDR is steadily increasing
- We have just 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
 - We have devised a logical structure allowing us to write 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

Conclusions (2)

Summing up

- The data oceans allow us to have homogeneous data from all studies in one place
 - Recoding selected dictionary-related variables is easy
- Data poolings are static snapshots of the (subset of) studies we need to support a certain analysis
- The importance and relevance of correct study metadata is very high, so a shift in approach is needed there to be able to fully exploit the benefits of a system like CDR
 - Quality of metadata is as critical as that of the data, if not more

Questions?



Pantaleo Nacci
Novartis Vaccines & Diagnostics Srl
Via Fiorentina, 1
53100 Siena, Italy
Email: pantaleo.nacci@novartis.com