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  - Ultimately, need to get rid of outdated, non-scalable processes & systems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

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>874</td>
<td>864</td>
<td>V59</td>
<td>V59P13E1</td>
</tr>
<tr>
<td>875</td>
<td>865</td>
<td>V59</td>
<td>V59P14E1</td>
</tr>
<tr>
<td>876</td>
<td>866</td>
<td>V102</td>
<td>V102_03</td>
</tr>
<tr>
<td>877</td>
<td>867</td>
<td>V102</td>
<td>V102_02E1</td>
</tr>
<tr>
<td>881</td>
<td>871</td>
<td>V72</td>
<td>V72_60</td>
</tr>
<tr>
<td>883</td>
<td>873</td>
<td>V102</td>
<td>V102_15</td>
</tr>
</tbody>
</table>
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)
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

NOVARTIS VACCINES
(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
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 lose 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
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
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’
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