



Going Against the Flow: Backmapping SDTM Data

Pantaleo Nacci, Head Statistical Safety & Epidemiology

PhUSE Annual Conference 2013

Brussels, 14 October 2013

Agenda



Introduction

The Clinical Data Repository

Moving Data to a New Standard

Developing the Program

Code Examples

Conclusions

Introduction

- NVD has embarked on a multi-year project to remap all available legacy study data from a proprietary standard to an industry-standard one
- While these legacy data are inspected and mapped to SDTM in a validated way, there are still urgent activities requiring the pooling of data
- In several instances most of the data involved in the analysis are still in the old standard
- This presentation talks about the experience gathered while developing a program to back-map data from the new CDISC-compliant standard to the previous one

Agenda



Introduction

The Clinical Data Repository

Moving Data to a New Standard

Developing the Program

Code Examples

Conclusions

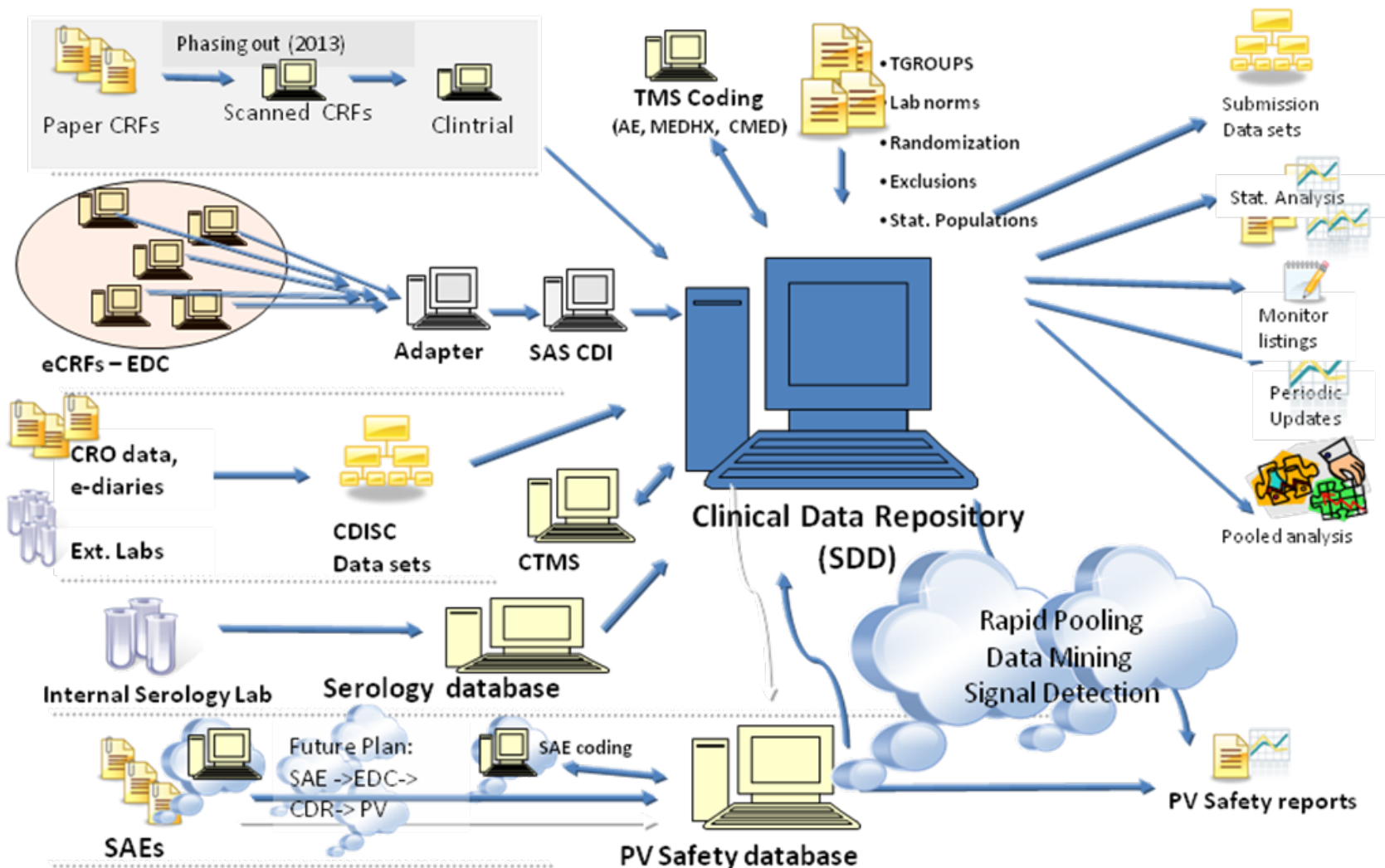
The Clinical Data Repository

A quick tour

- Clinical Data Repository (from now on simply CDR) is the name of the new NVD system for storing, managing and reporting on clinical studies, which went live in December 2011
- CDR has been developed to revolutionize NVD ability to:
 - Address complex health authority questions quickly and completely
 - Produce CDISC-compliant submissions
 - Review safety data in real-time
 - Mine our overall database for scientific and commercial queries
 - Improve overall productivity

An All-in-one System

Overall CDR structure



Agenda



Introduction

The Clinical Data Repository

Moving Data to a New Standard

Developing the Program

Code Examples

Conclusions

Moving Data to a New Standard

Pharma companies are on the move

- Many companies are already remapping legacy data to a common data standard or are planning to do it sooner rather than later
 - Usually one or more of the CDISC ones (SDTM/ADaM) are chosen
- In NVD the initial timelines to complete this exercise were both aggressive and optimistic,
 - The planned period during which data would be split between the two standards was extremely short
- In fact this period proved to be much longer
 - A combination of scarce experienced resources, competing activities and legacy data issues

NVD Data Oceans

Where all CDR data ultimately flow

- An automated process makes sure that all data from studies in CDR are added to one of two so-called 'oceans' as soon as they show up in specific study directories
 - Complete: contains data from all studies which are closed, inherently not very dynamic
 - Ongoing: includes data from running studies, dynamically rebuilt every night
- During the 'oceaning' process further harmonisation tasks are performed, e.g., all adverse events and medical history records are recoded to the latest version of MedDRA
- Global MedDRA terminology updates are also performed on the whole complete ocean twice per year

Using Oceans for Analysis

No more need to assemble data on the fly

- The initial plan stipulated that once all data for a certain product were in the oceans, they would be used as the only data source for any analysis spanning more than one study.
- The result of the delay mentioned earlier is a hybrid situation
 - For major products most of (but not always all) the data are in CDR, while several others are lingering behind at various stages of porting
- During this interim period several approaches have been devised by the programmers according to the task at hand
 - E.g., partial conversion of a subset of all data can be enough to run a DSUR/PSUR

A Proof of Concept

Another way of surviving with two data standards

- To keep options open and at the same time maintain my SAS skills up to speed, I developed a back-map program
 - Data collected in CDASH and then transformed into SDTM are ported back to the old, pre-CDISC Chiron/NVD standard
- The current version of the program is capable of recreating all the common panels for a total of 27 studies
- As of today it has not been formally validated, so that it cannot be used in production runs 'as is'
- An early version was validated by converting a subset of the back-mapped data to SDTM again
 - Some changes were needed, but ultimately it worked ok!

Agenda



Introduction

The Clinical Data Repository

Moving Data to a New Standard

Developing the Program

Code Examples

Conclusions

Developing the Program

Going back to the roots

- The exercise in itself was not dissimilar to what we had to do for the Legacy Data Conversion task
- First and foremost, a full understanding of the source data standard was needed, as well as of the target one
 - This time the source standard was CDISC SDTM, well documented by definition, so the real issues came from the other one

Identifying the Target Standard

Variability is always there

- The old standard was created in the mid-'90s by Chiron staff, and overall remained remarkably stable over time
 - Some variables went from numeric to character, or the other way round (e.g., month part of dates)
 - A few variables were used in different studies to store different information
 - Over time different variables were used to collect the same info
 - Several versions of non-standard panels were used
 - The implementation of EDC led to existing variables being sometimes defined with unexpected lengths
 - In early CDASH studies, whose data were remapped (or rather renamed) to the old standard within Clintrial, most variables ended up with a length of 200 by default

Pick and Choose

Only the best will do

- In the end I selected what I considered the best representation of the standard structure for each panel from different studies
 - E.g., DEMOG and CBP came from one study, ADVERSE and CMED from another, and so on
- Variable naming conventions were not always applied consistently, and that needed to be fixed too

Example: Inconsistent Naming Conventions

25	IRED_CRF	Character	10	\$10.	\$10.	INJ SITE STUDY VACCINE ERYTHEMA ON CRF
26	IREDUNIT	Character	2	\$2.	\$2.	INJ SITE STUDY VACCINE ERYTHEMA UNIT
27	IREDD	Numeric	8			INJ SITE STUDY VACCINE ERYTHEMA (MM)
28	IHARD_CR	Character	10	\$10.	\$10.	INJ SITE STUDY VACCINE INDURATION ON CRF
29	IHARD	Numeric	8			INJ SITE STUDY VACCINE INDURATION (MM)
30	IHARDUNI	Character	2	\$2.	\$2.	INJ SITE STUDY VACCINE INDURATION UNIT
31	CHILLS	Character	2	\$2.	\$2.	CHILLS
32	CHILLSW	Character	4			CHILLS
33	MALAISE	Character	2	\$2.	\$2.	MALAISE
34	MALAISEW	Character	4			MALAISE
35	MYALGIA	Character	2	\$2.	\$2.	MYALGIA
36	MYALGIAW	Character	4			MYALGIA
37	FATIGUE	Character	2	\$2.	\$2.	FATIGUE
38	FATIGUEW	Character	4			FATIGUE
39	ARTHRALG	Character	2	\$2.	\$2.	ARTHRALGIA
40	ARTHRALW	Character	4			ARTHRALGIA
41	QUICK	Character	1	\$1.	\$1.	QUICK BOX
42	QUICKW	Character	16			QUICK BOX
43	HEADACHE	Character	2	\$2.	\$2.	HEADACHE
44	HEADACHW	Character	4			HEADACHE
45	ECCH_CRF	Character	10	\$10.	\$10.	MEASURED ECCHYMOSIS ON CRF
46	ECCHUNIT	Character	2	\$2.	\$2.	ECCHYMOSIS UNIT
47	ECCHD	Numeric	8			ECCHYMOSIS IN MM
48	SXFEVER	Character	2	\$2.	\$2.	FEVER POST INJECTION
49	SXFEVERW	Character	7			FEVER POST INJECTION
50	SWEL_CRF	Character	10	\$10.	\$10.	SWELLING ON CRF
51	SWEL_UNI	Character	2	\$2.	\$2.	SWELLING UNIT
52	SWELD	Numeric	8			SWELLING DERIVED IN MM

The Other Side of the Coin

Is SDTM enough to backmap?

- Even if issues tend to be less pronounced, there is variability in our own CDASH/SDTM implementations too
 - Two SDTM versions, different in several subtle ways, were ultimately used, one for Legacy Data Conversion, the other one for CDR-originated studies
 - E.g., we realised that compliance phone call data should have been included in DS only after Phase 1 of the remapping was completed
- The other point which became clear after a while was that SDTM was not always sufficient to recreate the old data structures
 - SDTM has no allowance for administrative variables, so access to CDASH data was needed to get those (if collected!)

Agenda



Introduction

The Clinical Data Repository

Moving Data to a New Standard

Developing the Program

Code Examples

Conclusions

Checking for the Existence of XX/SUPPXX Pairs

- The first step is to merge supplemental domains back into the main ones: the logic used to check the existence of non-empty data sets was as follows:

```
%macro combine_supp(domain=);  
%if %sysfunc(exist(sdtm.&domain)) %then %do;  
    %let dsid = %sysfunc(open(sdtm.supp&domain));  
    %if &dsid %then %do;  
        %let isobs = %sysfunc(attrn(&dsid,any));  
        %let rc = %sysfunc(close(&dsid));  
    %end;  
    %else %let isobs = 0;  
    %if %sysfunc(exist(sdtm.supp&domain)) & &isobs=1 %then %do;  
        <etc.>  
    %end;  
%end;  
%mend;  
%combine_supp(domain=ae);  
...
```

Building Dates Using ISO 8601 Values

- Another generic macro dealt with obtaining dates and times from --DTC variables

```
%macro get_date(in=, out=);  
  if &in.dtc ^= '' then do;  
    &out.dy = input(scan(&in.dtc, 3, '-T'), 2.);  
    &out.mo = scan(&in.dtc, 2, '-');  
    &out.yr = input(scan(&in.dtc, 1, '-'), 4.);  
    &out.dt = mdy(max(input(&out.mo, 2.), 1), max(&out.dy, 1), &out.yr);  
    if &out.mo ^= '10' then &out.mo = compress(&out.mo, '0');  
    if index(&in.dtc, 'T') > 0 then &out.tm = input(scan(&in.dtc, 2, 'T'), time5.);  
  end;  
%mend;
```

Recoding Variable Values

- In the old standard many values were stored as codes in one variable, and as clear text in the corresponding –W one

```
proc format;
  ...
  value $vyn
    'N' = '2'
    'Y' = '1'
    'NA' = '3'
  ;
run;

data adverse;
  set ae;
  ...
  serious = put(aeser, $vyn.);
  seriousw = put(serious, $yesnos.);
  ...
run;
```

Finding Death Information

- There is no domain for death data in SDTM 3.1.2, so it was unclear where the relevant data might be found

```
data death_add1 (keep = prot ptno deathdy deathmo deathtyr deathdt deathtm causesp);  
  length prot $ 18  
         ptno $ 7;  
  set ae (where = (aesdth = 'Y'));  
  prot = studyid;  
  ptno = scan(usubjid, 2, '-');  
  %get_date(in=aest, out=death);  
  causesp = aeterm;  
run;  
  
data death_add2 (keep = prot ptno);  
  length prot $ 18  
         ptno $ 7;  
  set ds (where = (upcase(dsterm) = 'DEATH'));  
  prot = studyid;  
  ptno = scan(usubjid, 2, '-');  
run;
```

Solicited Adverse Events Data

- The structure of these data changed completely when the new EDC panels were designed, going from 'fat' to 'skinny'
- A lengthy routine was necessary to make the conversion process reasonably flexible and thus easily maintainable

```
%let sev_list = IPAIN ARTHRALG CHILLS FATIGUE HEADACHE DIARRHEA INAPPET MFEVER  
MALAISE MYALGIA NAUSEA PRURITIS SWEAT VOMIT RASH  
PTENDSV EATCHSV IRRITSV SLEEPSV;  
%let pre_list = ANALGESI ANAANTP ANAANTT;  
%let mea_list = ECCH IHARD IRED SWEL;  
  
data postinj_  
length  
%do i = 1 %to %sysfunc(countw(&sev_list));  
%scan(&sev_list, &i) $ 2  
%upcase(%sysfunc(compress(%sysfunc(subpad(%scan(&sev_list, &i), 1, 7))))w) $ 8  
%end;
```

Solicited Adverse Events Data (2)

- As an additional complication, data once collected in the same panel are now split over different domains, e.g., body temperature in VS

```
data vs_postinj;  
  set vs (where = (scan(vstpt, 1) = 'DAY' & vsstat ^= 'NOT DONE'));  
  select(vstestcd);  
    when('TEMP') do;  
      temp = input(vsorres, best8.); tempunit = upcase(vsorresu);  
      tempuniw = tempunit; tempd = round(vsstresn, 0.1); temploc = vsloc;  
      visnum = visitnum;  
    end;  
  end;  
run;
```

Agenda



Introduction

The Clinical Data Repository

Moving Data to a New Standard

Developing the Program

Code Examples

Conclusions

Conclusions

Your mileage may vary

- The benefits of having all available study data in one place and in a common, well defined format are many and covering a wide range of applications
 - To be able to reap them the remapping period should be kept as short as possible
- Stopping or even slowing down the remapping exercise mid-way is a dangerous proposition
 - There is a real risk of creating more confusion rather than a clearer situation
- Having to guess on which of two platforms an analysis will run with the least problems is not a good thing

Conclusions (2)

- My backmapping program was born as a proof of concept and is still a work in progress
 - Almost every time a new study goes live in CDR a new slight variation of something requires a change in the program
 - It might be just a new LBTESTCD value, easily addressable, or the appearance of a brand new domain
- Until now I managed to maintain only one version of the program, dealing with all studies
 - Keeping it in a validated status would be difficult
 - The one time it was done, the fastest way to validate it we could think of was to have another programmer convert again the data to SDTM, and then check that data were the same at the end of the complete circle

Any Questions?

