Legacy Data Conversion: A Journey of Discovery

Pantaleo Nacci, Head Statistical Reporting

PhUSE Annual Conference 2012

Budapest, 16 October 2012

NOVARTIS
VACCINES

Agenda

Introduction

The Clinical Data Repository

The Legacy Data Conversion Sub-project

Helping Programmers Do Their Job

Learn to Know Your Metadata

Conclusions

NOVARTIS
VACCINES

Introduction

- NVD has been working for almost three years on a very challenging initiative to set up a brand new validated environment for clinical trial data collection, management, analysis, and reporting, as well as to remap all existing study data
- In Berlin I gave a presentation about a previous exercise in data pooling, which remains one of the main business drivers behind the new porting
- This presentation will talk about the experience we had in Novartis Vaccines while converting (part of) our legacy data from an old, proprietary standard (more than one, actually) to a modern, industry-wide one

Agenda

Introduction

The Clinical Data Repository

The Legacy Data Conversion Sub-project

Helping Programmers Do Their Job

Learn to Know Your Metadata

Conclusions

NOVARTIS
VACCINES

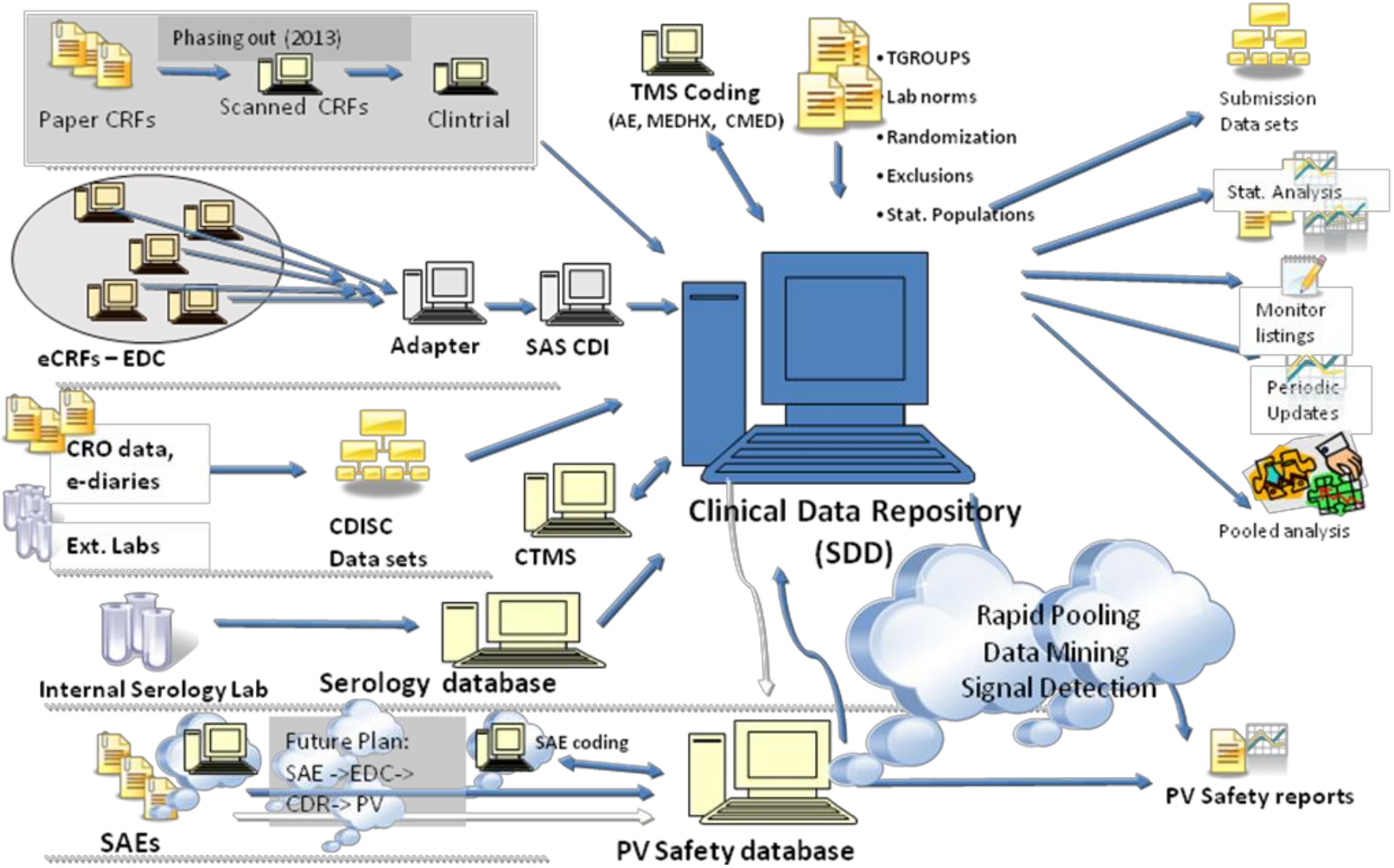
The Clinical Data Repository

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.
- CDR has been developed to revolutionize our 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 in Global Clinical Research & Development

The eCloud

Overall CDR structure



NOVARTIS

Agenda

Introduction

The Clinical Data Repository

The Legacy Data Conversion Sub-project

Helping Programmers Do Their Job

Learn to Know Your Metadata

Conclusions

NOVARTIS VACCINES

The Legacy Data Conversion Sub-project

The story so far

- Novartis (previously Chiron) Vaccines has accumulated data from clinical trials for two decades, the earliest ones dating back to 1992, and spanning over a considerable number of vaccines (influenza, meningitis, rabies, TBE, Hib, HIV, TDaP, HBV, etc.) as well as several adjuvants and preservatives
- The need to pool data from multiple trials surfaced dramatically during the recent A/H1N1 pandemic, when companies received from Regulatory Agencies multiple requests for retrospective safety analyses of all data collected in selected trials, always with tight turnaround times
- To be able to pool efficiently and effectively, all data (and metadata) need to share the same structures in the first place

Early Choices

What needs to be decided before the start

- List of available studies
  - In our case ‚legacy' was any study not using CDASH eCRFs, thus including recent studies using the old standard and still running
- Which (subset of) studies to port first
  - It would be advisable to have a pilot phase, so that unplanned issues in the process are identified while activities are still on a manageable scale, but in the end we skipped it (and paid the price for doing so)
  - In our case the pool to choose from included roughly 400 studies
  - Top management was asked to come up with criteria allowing us to identify the top priority ones, and so they did
  - The selection was based on completed studies in high priority projects or containing a certain chemical entity, for a final total of 153

Data Inspection

Getting to know the data

- The selected studies could be split in pre- and post-Clintrial adoption:
  - Clintrial studies had been managed using Clintrial 3 or 4, and were based on minor variations of the old non-CDISC standard inherited from Chiron, well known by programming group
  - Pre-Clintrial ones on the contrary were managed by external CROs in the first half of the 90's, so that knowledge on them was extremely limited
- Clintrial 4 is still used today, and IT managed to restore an instance of Clintrial 3, so we could download fresh versions of all those study data
- For the other studies we had to rely on whatever data were available on our existing SAS server

Raw Data Are Not Enough

Pitfalls

- Once we had the Clintrial data we compared them to what was used for the original analysis, and all discrepancies were documented and investigated
- We then tried to locate the original study documentation (protocols, amendments, CRFs, CSRs)
- In some cases we failed, completely or in part, as expected especially for older studies
- In the early days amended protocols were not re-issued as we do now, so in at least one case we found out that we were missing something only by looking at the data

Establishing what a Variable Contains

A risk linked to incomplete documentation

- We had to guess the meaning of variables by comparing them with those found in similar studies
- This way proved mostly ok, but in a few cases the dangers of such an approach were evident, like in the example on the right

RSERUMD	RSGPT	RSGPTCS	COMMENT
11/25/92	68		C

SGPTDT	RSGPT	VSGPT	COMMENT
	2		
	2		
	2		
	2		
	2		
11/25/92	1	46	C
02/01/93	1	57	

Agenda

Introduction

The Clinical Data Repository

The Legacy Data Conversion Sub-project

Helping Programmers Do Their Job

Learn to Know Your Metadata

Conclusions

NOVARTIS
VACCINES

Helping Programmers Do Their Job

And increase overall homogeneity at the same time

- To facilitate the work of the programmers, mapping datasets were centrally developed for each domain
- They contained a variable number of records for each study to be remapped, depending on the domain
- A blank cell would mean that no legacy variable had been identified as containing that particular information
- Further investigation led in some cases to the identification of another variable, but in others the info had simply not been collected
- A list of all unmapped variables and panels was created

Example of a Mapping Data Set

	DOMAIN	VAC	LBN	DSN	LOC	VERS	STUDYID	USUBJID	EXLOC	VISIT	VISITNUM	EXGRPID	EXSTDTC	EXSTDTC2	EXENDTC
23	EX	HIV	V24P1	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V24\V24P1\	1	PROT	PTNO	SITEW	VISIT	VISIT	INJNO	STARTDT	STARTTM	STARTDT
24	EX	FLUN	V25P1	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V25\V25P1\	1	PROT	PTNO		VISIT	VISIT	INJNO	STARTDT	STARTTM	STARTDT
25	EX	HIV	V26V6P1	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V26V6\V26V6P1\	1	PROT	PTNO	SITEW	VISIT	VISIT	INJNO	STARTDT	STARTTM	STARTDT
26	EX	HCV	V35P1	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V35\V35P1\	1	PROT	PTNO	SITEW	VISIT	VISIT	INJNO	STARTDT	STARTTM	STARTDT
27	EX	HBV	V42P1	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V42\V42P1\	1	PROT	PTNO	SITEW	VISIT	VISIT	INJNO	STARTDT	STARTTM	STARTDT
28	EX	HBV	V42P2	DOSING	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V42\V42P2\	1	PROT	PTNO	SITEW		VISNUM		STARTDT	STARTTM	STARTDT
29	EX	HIV	V46P1	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V46\V46P1\	5	PROTOCOL	SUBJECT	INJSITE	VISIT	VISIT		DATET	INJTIME	DATET
30	EX	HIV	V4P3	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V4\V4P3\	1	PROT	PTNO	SITEW	VISIT	VISIT	INJNO	IMMUNDT	IMMUNTM	IMMUNDT
31	EX	FLUN	V57P1	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V57\V57P1\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
32	EX	MENACWY	V59P1	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P1\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
33	EX	MENACWY	V59P10	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P10\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
34	EX	MENACWY	V59P11	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P11\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
35	EX	MENACWY	V59P13	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P13\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
36	EX	MENACWY	V59P14	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P14\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
37	EX	MENACWY	V59P16	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P16\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
38	EX	MENACWY	V59P17	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P17\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
39	EX	MENACWY	V59P18	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P18\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
40	EX	MENACWY	V59P1E1	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P1E1\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
41	EX	MENACWY	V59P2	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P2\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
42	EX	MENACWY	V59P20	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P20\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
43	EX	MENACWY	V59P21	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P21\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
44	EX	MENACWY	V59P22	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P22\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
45	EX	MENACWY	V59P2E1	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P2E1\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
46	EX	MENACWY	V59P2E2	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P2E2\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
47	EX	MENACWY	V59P3	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P3\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
48	EX	MENACWY	V59P4	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P4\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
49	EX	MENACWY	V59P5	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P5\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
50	EX	MENACWY	V59P5E1	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P5E1\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
51	EX	MENACWY	V59P6	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P6\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
52	EX	MENACWY	V59P7	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P7\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
53	EX	MENACWY	V59P8	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P8\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
54	EX	MENACWY	V59P9	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V59\V59P9\	8	PROT	PTNO	SITEW	VISNUM	VISNUM	INJNO	STARTDT	STARTTM	STARTDT
55	EX	HSV	V5P1	VACCINE	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V5\V5P1\	5	PROTOCOL	SUBJECT	INJSITE	VISIT	VISIT		DATE		DATE
56	EX	HSV	V5P10	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V5\V5P10\	6	PROTOCOL	SUBJECT	INJSITE	VISIT	VISIT		DATET	INJTIME	DATET
57	EX	HSV	V5P11	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V5\V5P11\	6	PROTOCOL	SUBJECT	INJSITE	VISIT	VISIT		DATET	INJTIME	DATET
58	EX	HSV	V5P12	IMMUN	E:\CDR\SOURCE_DATA\PHASE1\DEVEL\OUTPUT\V5\V5P12\	6	PROTOCOL	SUBJECT	INJSITE	VISIT	VISIT		DATET	INJTIME	DATET

Agenda

Introduction

The Clinical Data Repository

The Legacy Data Conversion Sub-project

Helping Programmers Do Their Job

Learn to Know Your Metadata

Conclusions

NOVARTIS
VACCINES

Learn to Know Your Metadata

The real key to using your data

- In my opinion the addition of the T-domains is among the major improvements in CDISC SDTM 3.1.2
- Since among our goals was also the ability to slice and dice all available data, this aspect was critical
- The incomplete status of the original documentation for some early studies made their completion problematic
- To complement the T-domains we designed two custom metadata domains, actually rather lookup tables
- They were designed to allow us to establish exactly to which chemical or biological components each subject has been exposed to in the course of a clinical study

Vaccine Component (VC)

The building blocks

- The first one lists all relevant components found in any vaccine or drug ever used in our studies, like thiomersal

DOMAIN	COMPCD	COMPTY	COMPSNM	COMPLNM	COMPDS	COMPUN
VC	C0067	OTHER	THIOMERSAL	Thiomersal	0.05	mg
VC	C0068	OTHER	THIOMERSAL	Thiomersal traces		
VC	C0069	ADJUVANT	LTK63	LTK63 + SMVB	3	mcg
VC	C0070	ADJUVANT	LTK63	LTK63 + SMVB	10	mcg
VC	C0071	ADJUVANT	LTK63	LTK63 + SMVB	30	mcg
VC	C0072	ADJUVANT	LTK63	LTK63	30	mcg
VC	C0073	ANTIGEN	FLU_H3N2	A/Moscow/10/99	7.5	mcg
VC	C0074	ANTIGEN	FLU_B	B/Hong Kong/330/2001	7.5	mcg
VC	C0075	ANTIGEN	FLU_H1N1	A/New Caledonia/20/99	7.5	mcg
VC	C0076	ANTIGEN	FLU_H3N2	A/Wisconsin/67/2005	7.5	mcg
VC	C0077	ANTIGEN	FLU_H1N1	A/Solomon Islands/3/2006	7.5	mcg
VC	C0078	ANTIGEN	FLU_B	B/Brisbane/60/2008	7.5	mcg
VC	C0079	ANTIGEN	MEN_C	MenC-CRM197, lyophilised	10	mcg
VC	C0080	ADJUVANT	ALUM	Aluminium hydroxyde		
VC	C0081	ANTIGEN	TBE	TBE, K23 strain	0.75	mcg
VC	C0082	OTHER	POLYGELINE	Polygeline		
VC	C0083	ANTIGEN	FLU_B	B/Beijing/184/93	7.5	mcg
VC	C0084	ANTIGEN	CMV_GB	Glycoprotein B	5	mcg
VC	C0085	ANTIGEN	CMV_GB	Glycoprotein B	30	mcg
VC	C0086	ANTIGEN	CMV_GB	Glycoprotein B	100	mcg

Vaccine Formulation (VF) and Its Use in T-domains

The link to understanding exposure

- The second uses a subset of records from VC to identify what was used in formulating a vaccine or drug
- VF records are then used in ELEMENT as needed
- A simple SAS program can then, e.g., retrieve the list of all entries in VF containing a VC code of interest

DOMAIN	FORMCD	FORMSNM	FORMLNM	FORMCMP	FORMDS	FORMUN	GENERIC
VF	F0019	FLUAD	Fluad Northern Hemisphere 2002/03 w/ thiomersal	C0004,C0014,C0025,C0043,C0067	0.5	mL	MF59-eTIV
VF	F0020	FLUAD	Fluad Northern Hemisphere 2002/03 w/ thiomersal traces	C0004,C0014,C0025,C0043,C0068	0.5	mL	MF59-eTIV
VF	F0021	FLUAD	Fluad Northern Hemisphere 2002/03	C0004,C0014,C0025,C0043	0.5	mL	MF59-eTIV
VF	F0022	FLUAD	Fluad Southern Hemisphere 2003	C0004,C0014,C0025,C0043	0.5	mL	MF59-eTIV
VF	F0023	FLUAD	Fluad Northern Hemisphere 2003/04	C0004,C0014,C0025,C0043	0.5	mL	MF59-eTIV
VF	F0024	FLUAD	Fluad Southern Hemisphere 2004	C0004,C0015,C0025,C0043	0.5	mL	MF59-eTIV
VF	F0025	FLUAD	Fluad Northern Hemisphere 2004/05	C0004,C0015,C0026,C0043	0.5	mL	MF59-eTIV

STUDYID	DOMAIN	ARMCD	ARM	TAETORD	ETCD	ELEMENT	TABRANCH
V70P1	TA	A	FLUAD_N	1	SCR	SCREENING	RANDOMIZED TO A
V70P1	TA	A	FLUAD_N	2	FLUAD21	FLUAD0021	
V70P1	TA	B	FLUAD_O	1	SCR	SCREENING	RANDOMIZED TO B
V70P1	TA	B	FLUAD_O	2	FLUAD20	FLUAD0020	

Agenda

Introduction

The Clinical Data Repository

The Legacy Data Conversion Sub-project

Helping Programmers Do Their Job

Learn to Know Your Metadata

Conclusions

NOVARTIS VACCINES

Conclusions

There is still a lot to do!

- Converting legacy data can easily prove a nightmare, and will be at least a very challenging task
- There are other aspects we didn't touch (e.g., recoding of verbatims, consistency checks, validation strategy, programming oversight, project management, populating the T-domains) but anyone embarking in a similar exercise will need to think about them too
- Decisions taken too early with incomplete or draft information can easily lead to a blind alley
- Parallel development of logically sequential processes should be avoided

Conclusions (2)

- Full understanding of the data to be remapped is critical, including access to enough documentation to allow clear identification of all variables
  - Working on a ‘best guess’ basis will always entail a certain level of risk
- Data from studies run in the last century contain on average more mistakes than we are used to these days
- Documentation and validation activities require a lot of time and attention, and it’s difficult to find programmers who are good and enjoy doing it
- Keep your QA people in the loop at all times, they are there to help, not to hinder

Conclusions (3)

- Document all decisions and the reasoning behind them
  - It will be easier to understand clearly why a certain route was taken and defend it later
- Manage expectations of top management carefully, any task will always take longer than you expect
- Be aware of the ‘Big Red Button’ syndrome: all non-programmers believe that we have one of them somewhere, so that for us it’s just a question of pushing it to obtain the desired results

References

- CDISC: http://www.cdisc.org
- FDA: http://www.fda.gov/BiologicsBloodVaccines/DevelopmentApprovalProcess/ucm209137.htm

Question time

“Are you being served?”