Less is more - A visionary View on the Future of CDISC Standards

Presented by FH-Prof.Dr. Jozef Aerts
University of Applied Sciences FH Joanneum
Graz, Austria

FH | JOANNEUM
University of Applied Sciences

XML 4Pharma

Strength *through Collaboration*

Slides at: www.XML4Pharma.com/slides/

CDISC

Imagine …

- You want to buy a english book on Amazon
- So you load the file "books_en.xpt" (27 106 records)
  - => books_1_en.xpt, books_2_en.xpt, …
- Description of the content < 200 characters, title < 40 characters, only US-ASCII
  - => SUPP_books_1_en.xpt
- Comments about the book
  - => CO.xpt

Imagine …

- You found your book an want to know whether there is a movie about it
  - => RELREC.xpt
  - => Movies_n_en.xpt

- Would you still by a book on Amazon.com if this were reality?

This is SDTM and the FDA review process…

Is SDTM a database?
(and if so, it is a good one?)

- Lots of redundant information
  - Leads to errors and decreased data quality (more is worse …)

- Many derived variables (--DY, --EPOCH)
  - Submissions with >50% incorrect --DY values (go unnoticed in validation tools)
- Ambiguous information
  - LBTESTCD + LBCAT + LBMETHOD does not uniquely identify a lab test

http://cdiscguru.blogspot.com/2012/07/is-sdtm-database-design-and-if-so-is-it.html

The experiment … A real dataset

- Remove all --DY variables and EPOCH (obs.domains)
- Remove RFXSTDTC, RFXENDTC, ARM, ACTARM, … (DM)
- Remove --TEST, VISIT (name) (obs.domains)
- Remove LBTESTCD, LBTEST, LBCAT, LBSCAT, LBMETHOD, but keep LBLOINC (LB)
- Slightly change the variable label to better explain what it is about

Does this submission make it through P21 and FDA Datafit?

Although the data quality is potentially higher …

Experiment outcome - Inspecting the data …

DOMAIN	USUBJID	SUBJID	RFSTDTC	RFENDTC
DM	01-701-1015	1015	2014-01-02	2014-07-02
DM	01-701-1023	1023	2012-08-05	2012-09-0a
DM	01-701-1028	1028	2013-07-19	2014-01-14
DM	01-701-1033	1033	2014-03-18	2014-04-14
DM	01-701-1034	1034	2014-07-01	2014-12-30
DM	01-701-			
DM	01-701-			
DM	01-701-			
DM	01-701-1111	1111	2012-09-07	2012-09-17

01-701-1033 (USUBJID)
First date of study treatment exposure = Tue Mar 18 2014
Last date of study treatment exposure = Mon Mar 31 2014

COUNTRY	DMDTC
USA	2013-12-26
USA	2012-07-22
USA	2013-07-11
USA	2014-03-10
USA	
USA	

2013-07-11 (DMDTC)
DMDY = -8

VS

USUBJID	VSSEQ	VSTESTCD	VSPOS	VSO
01-701-1015	7	DIABP	SUPINE	56
01-701-1015	8	DIABP	STANDING	51
01-701-1015	9	DIABP	STANDING	61
01-701-1015	10	DIABP		
01-701-1015	11	DIABP		
01-701-1015	12	DIABP		

DIABP (VSTESTCD)
Diastolic Blood Pressure
NCI: C25299

VISITNUM	VSDTC	VSTPT
6	2014-02-01	AFTER LYIN
6	2014-02-01	AFTER STAN
6	2014-02-01	AFTER STAN
7		AFTER LYIN
7		AFTER STAN
7		AFTER STAN

6 (VISITNUM)
Visit: AMBUL ECG REMOVAL
Planned Study Day of Visit: 30

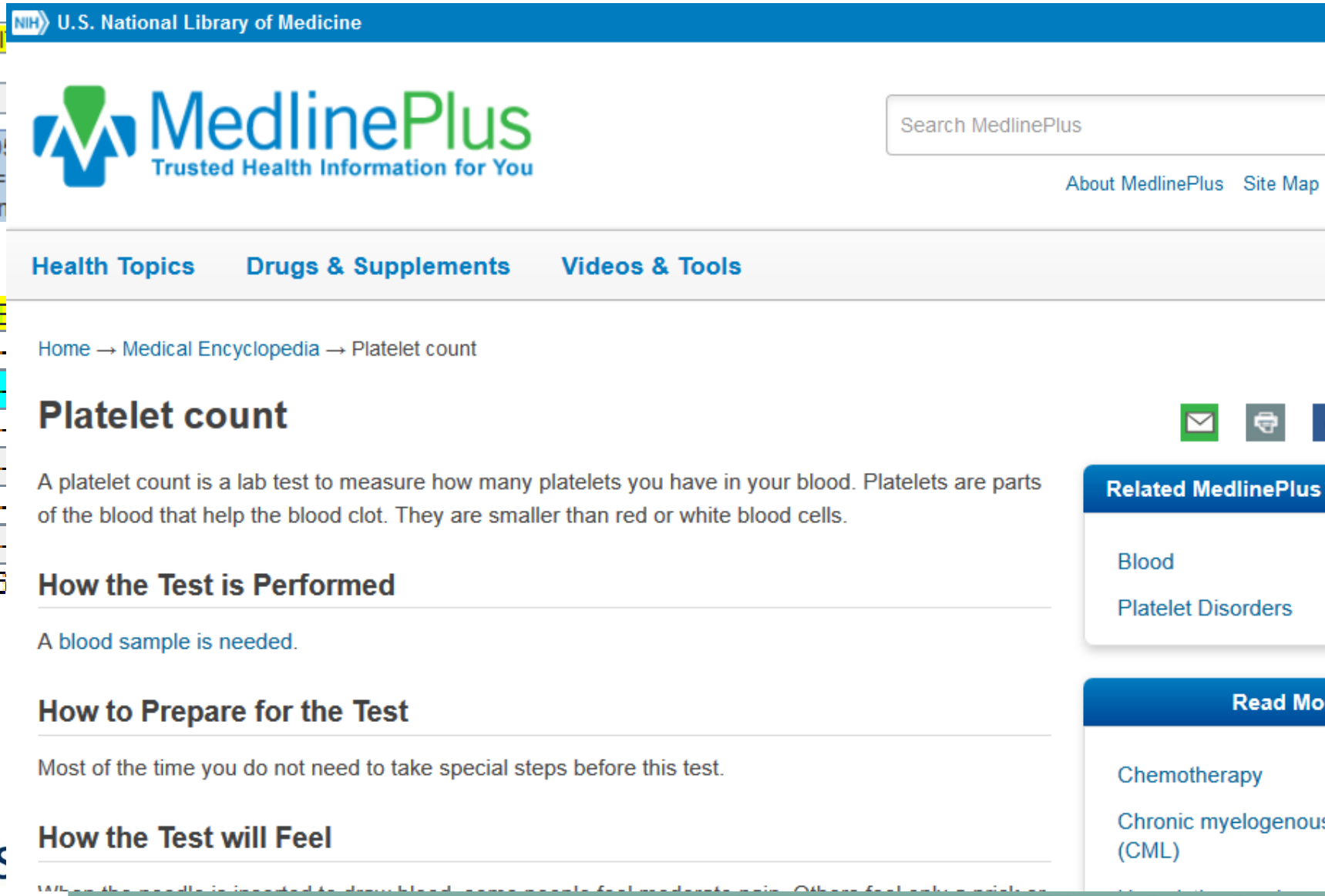
VISITNUM	LBDTC
9	2014-03-26T15:15
10	2014-05-07T11:21
11	2014-05-21T10:58

2014-05-07T11:21 (LBDTC)
LBDY = 125
Element = PBO (Placebo) - Epoch = Treatment

Experiment outcome - Inspecting the data …

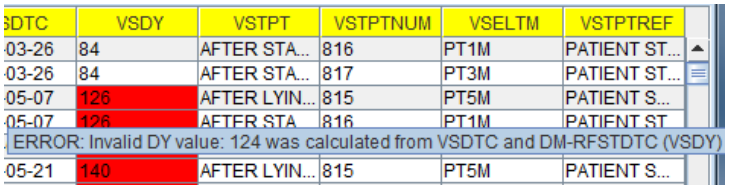
VISI
9
10
11
2014-05
LBDY =
Elemen

LE
26515-
26515-
26515-
26515-
26515-
26515-
2888-6



Experiment outcome

- No information was lossed, although
  - ~30% of the variables was removed
  - Including "FDA required" ones



SDTC	VSDY	VSTPT	VSTPTNUM	VSELTM	VSTPTREF
03-26	84	AFTER STA...	816	PT1M	PATIENT ST...
03-26	84	AFTER STA...	817	PT3M	PATIENT ST...
05-07	126	AFTER LYIN...	815	PT5M	PATIENT S...
05-07	126	AFTER STA...	816	PT1M	PATIENT ST...

ERROR: Invalid DY value: 124 was calculated from VSDTC and DM-RFSTDTC (VSDY)

05-21	140	AFTER LYIN...	815	PT5M	PATIENT S...

- Data quality improved
  - In the original dataset 40% of the --DY variable values was wrong
  - But nobody noticed (even the P21 validator didn't)
  - We now have unique lab test codes (through LOINC)
  - If we had used UCUM instead of [UNIT] we could have done unit conversions on-the-fly

Less is more!

So … what happened?

- The tool did cross-dataset lookup
  - And did simple calculations
- The tool did cross-dataset validation
- The tool used publicly available RESTful web services
  - From XML4Pharma (e-SHARE content) ~ 35 services
  - From the National Library of Medicine (NLM)
  - And soon be delivered by the SHARE API

www.XML4PharmaServer.com/WebServices

SDTM Variables that could be removed from LB

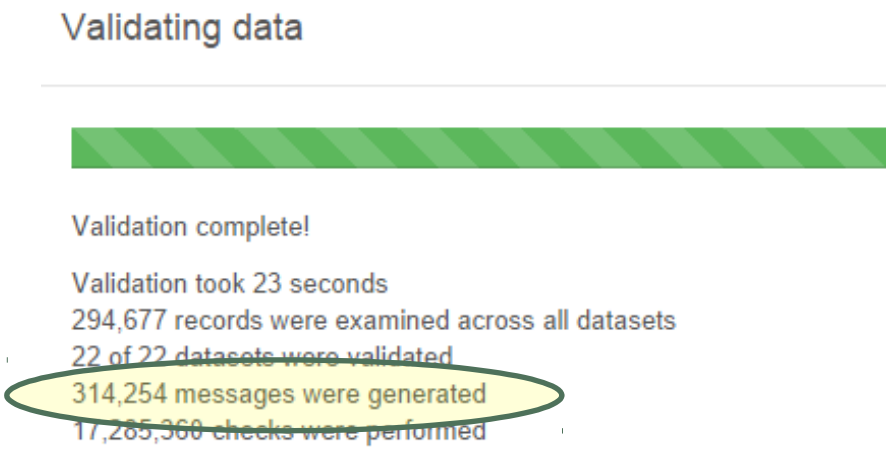
Variable	Reason
LBTEST	1:1 relation with LBTESTCD, provided by define.xml
LBCAT, LBSCAT	When LBTESTCD is provided as LOINC code Lookup possible through WS
[LBSTRESN]	If UCUM units used for LBORRESU, automatically calculated
[LBSTRNRLO, LBSTNRHI]	If UCUM units used for LBORRESU, automatically calculated from LBORNRLO and LBORNRHI
LBSPEC, LBMETHOD	Already provided by the LOINC code
[VISIT]	1:1 relation with VISITNUM when planned visit - lookup in TV
VISITDY	Planned Visit day - lookup in TV
LBDY, EPOCH	Derived - can easily be done by the tool

If we had the courage to remove these variables from SDTM (and use LOINC and UCUM), data quality would considerably improve

Less is more: Validation Rules and Data Quality

- CDISC standards were developed in order to get:
  - Easy of exchange, ease of review
  - Better data quality, comparable data
  - Bringing new therapies to the patient faster

- Validation Rules (define-XML, SDTM/SEND, ADaM) were introduced to ensure quality



Data Quality and Validation Rules

- Currently, the existing validation rules implementations:
  - Retard the review process
  - Overinterpretation, false positives, …
  - Often decrease data quality
  - => Tendency to "fix" validation issues …

IDVAR	IDVARVAL	QNAM
LBSEQ	1	LBTMSHI
LBSEQ	2	LBTMSHI
LBSEQ	3	LBTMSHI

SUPPLB					
SD0077	FDAC074	Invalid referenced record	Error	64403
SD1082	FDAC036	Variable length is too long for actual data	Error	6
SD1021	FDAC216	Unexpected character value in IDVARVAL variable	Warning	64403

3763	SUPPLB	1		IDVARVAL	[7 spaces]1	SD1021	FDAC216	Unexpected character value in IDVARVAL variable
3764	SUPPLB	2		IDVARVAL	[7 spaces]2	SD1021	FDAC216	Unexpected character value in IDVARVAL variable
3765	SUPPLB	3		IDVARVAL	[7 spaces]3	SD1021	FDAC216	Unexpected character value in IDVARVAL variable
3766	SUPPLB	4		IDVARVAL	[7 spaces]5	SD1021	FDAC216	Unexpected character value in IDVARVAL variable

Data Quality - example: labels

- Experiment: put a dot at the end of each SDTM label in your submission
  - You will get hundreds of errors

- Experiment: replace "Std Format" by "Std. Format" in a label
  - It might take you hours to find out what's wrong

- I have seen submission delays of weeks due to such issues
  - Write many many pages of "false positive explanations" in your Reviewer's Guide

- Sometimes you can better explain what the variable is about by slightly changing the variable label
  - But you are even not allowed …

Risk assessment instead of strict rules

- Strict rules are for dumb people …
  - "switch off your brain …"
- Risk assessment is a much better way
- Example: Labels: Equality Number for strings *

SDTM variable	Expected Label	Actual Label	Equality Number
LBTESTCD	Lab Test or Examination Name	Laboratory Test or Examination Name	0.80
MBSTRESC	Character Result/Finding in Std Format	Character Result/Finding in Std. Format	0.97
TADTC	Date/Time of Accountability Assessment	Date/Time of Drug Accountability Assessment	0.88
MBSTRESC	Character Result/Finding in Std Format	The quick brown fox jumps over the lazy dog	0.12

- But a semantic risk assessment would even be better

* Based on Levenhstein Algorithm

What can we do better?

- Make validation rules more flexible (and "smart")
  - Risk assessment instead of pass/fail rules

- Use modern technology, like:

- RESTful web services
  - Corrections and bug fixes within hours - not "next release" …

- Smart Viewers and Tools
  - Cross-domain lookups

- FDA: Data in databases, not in files
  - Files are only temporary means of transport

Flexible and smart validation rules

- Validation rules should be human-readable and machine-executable
- No "over-interpretation" of the standard
- Anthony Chow, Sam Hume and I are currently looking into such "rule standards/implementations"
  - OMG OCL
  - HL7 Gello
  - XQuery

- CDISC might publish "Reference Implementations"
  - Implementation of vendors should give identical results

XQuery implementation of ADaM validation rule

```
(: Rule ADaM Validation 1.3 rule 2: Any ADaM variable whose name is the same as an SDTM variable must be a copy of
must not be modified :)
(: Any ADaM variable whose name is the same as an SDTM
variable must be a copy of the SDTM variable, and its label and values must not be modified :)
xquery version "3.0";
declare namespace def = "http://www.cdisc.org/ns/def/v2.0";
declare namespace odm="http://www.cdisc.org/ns/odm/v1.3";
declare namespace data="http://www.cdisc.org/ns/Dataset-XML/v1.0";
declare namespace xlink="http://www.w3.org/1999/xlink";
let $base := '/db/fda_submissions/ADaM_cdiscpilot01/'
let $define := 'define_2_0.xml'
let $sdtmdefine := 'define_2_0_SDTM.xml'
let $definedoc := doc(concat($base,$define))
let $sdtmdefinedoc := doc(concat($base,$sdtmdefine))
(: iterate over all ADaM variables (as define by ItemDefs), find the corresponding one in the SDTM define.xml and
for $itemdef in $definedoc//odm:ItemDef
    let $name := $itemdef/@Name
    let $oid := $itemdef/@OID
    let $parentelement := name($definedoc//odm:ItemRef[@ItemOID=$oid]/..)
    let $label := $itemdef/odm:Description
    (: and the corresponding ItemDef in the SDTM :)
    for $sdtmitemdef in $sdtmdefinedoc//odm:ItemDef[@Name=$name]
        let $sdtmlabel := $sdtmitemdef/odm:Description
        let $sdtmoid := $sdtmitemdef/@OID
        let $sdtmparentelement := name($sdtmdefinedoc//odm:ItemRef[@ItemOID=$sdtmoid]/..)
        (: and compare both - the must be equal - exclude Valuelist level ItemDefs :)
        where not($label = $sdtmlabel) and $parentelement = 'ItemGroupDef'
            and $sdtmparentelement = 'ItemGroupDef'
        return <error rule="ADaM_v1-3_validation_rule_2" rulelastupdate="2015-09-18">ADaM variable {data($name)} h
corresponding SDTM variable for which the label '{data($sdtmlabel)}' was found</error>
```

ADaM Validation Rule "Any ADaM variable whose name is the same as an SDTM variable must be a copy of the SDTM variable…"

XQuery + RESTful WS = Vendor-neutral Validation

- No more false positives
- Always up-to-date (retrieve them from SHARE)
- Human-readable, machine-executable rules by CDISC

DSDECOD	DSCAT	VISITNUM
COMPLETED	DISPOSITION EVENT	13
FINAL LAB VISIT	OTHER EVENT	13
ADVERSE EVENT	DISPOSITION EVENT	5
FINAL LAB VISIT		
FINAL RETRIEVA		
COMPLETED		
FINAL LAB VISIT		

WARNING: Value 'FINAL LAB VISIT' is not a valid value of the extensible codelist NCOMPLT and has not been declared as an extended value in the define.xml (DSDECOD)

```
14
15 declare variable $base external;
16 declare variable $define external;
17
18 (: function to check whether a value is one of the given list :)
19 declare function functx:is-value-in-sequence
20   ( $value as xs:anyAtomicType? ,
21     $seq as xs:anyAtomicType* )  as xs:boolean {
22    $value = $seq
23  } ;
24
25 (: get the OID of the DSDECOD variable :)
26 let $dsdecodoid := doc(concat($base,$define))//odm:ItemDef[@Name='DSDECOD']/@OID
27 (: Get the location of the DS dataset :)
28 let $dsitemgroupdef := doc(concat($base,$define))//odm:ItemGroupDef[@Name='DS']
29 let $dsdatasetname := $dsitemgroupdef/def:leaf/@xlink:href
30 let $dsdatasetpath := concat($base,$dsdatasetname)
31
32 (: get the extended values from the codelist, and a list of allowed values from a web service -
33  this reduces the number of web service calls to a single one :)
34 let $dsdecodcodelistoid := doc(concat($base,$define))//odm:ItemDef[@Name='DSDECOD']/odm:CodeListRef/@CodeListOID
35 (:  let $dsdecodvaluesfromdefine := doc(concat($base,$define))//odm:CodeList[@OID=$dsdecodcodelistoid]/odm:*[not(@def:Extendedvalue='Yes')],
36 let $dsdecodextendedvalues := doc(concat($base,$define))//odm:CodeList[@OID=$dsdecodcodelistoid]/odm:*[@def:ExtendedValue='Yes']/@CodedValue
37 let $webservice := 'http://www.xml4pharmaserver.com:8080/CDISCCTService/rest/CodedValuesFromCodeListName/NCOMPLT'
38 (: the webservice returns an XML document with the structure /XML4PharmaServerWebServiceResponse/Response and 'CodedValue' elements -
```

Use of RESTful Web Services

- In Healthcare, HL7-FHIR is revolutionizing interoperability, due to ist implementation of RESTful Web Services
- None of the CDISC standards currently supports web services
- Next generation of CDISC standards needs to support RESTful web services

LBLOINC	LBORRES	LBORRESU	LBORNRLO
26515-7	263	THOU/uL	130
26515-7	252	THOU/uL	130
26515-7	268	THOU/uL	130
26515-7			
26515-7			
26515-7			
2888-6			

26515-7 (LBLOINC)
LOINC Name: Platelets:NCnc:Pt:Bld:Qn
LOINC Common Name: Platelets [#/volume] in Blood
Example UCUM Units: 10*3/uL

Next generation of CDISC Standards and RESTful Web Services

Provisional…

RESTful web service in ODM

```
<StudyEventDef OID="SE.UNPLANNED" Name="Unplanned Visit" Type="Unscheduled">
    <!-- HTTP GET request to SHARE to get the CDASH v1.1 AE form -->
    <FormRef Reference="http://share.cdisc.org/rest/CDASH/v1.1/Form.AE"/>
    <!-- HTTP GET request to SHARE to get the CDASH v1.1 CM form -->
    <FormRef Reference="http://share.cdisc.org/rest/CDASH/v1.1/Form.CM"/>
</StudyEventDef>
```

Less is more!

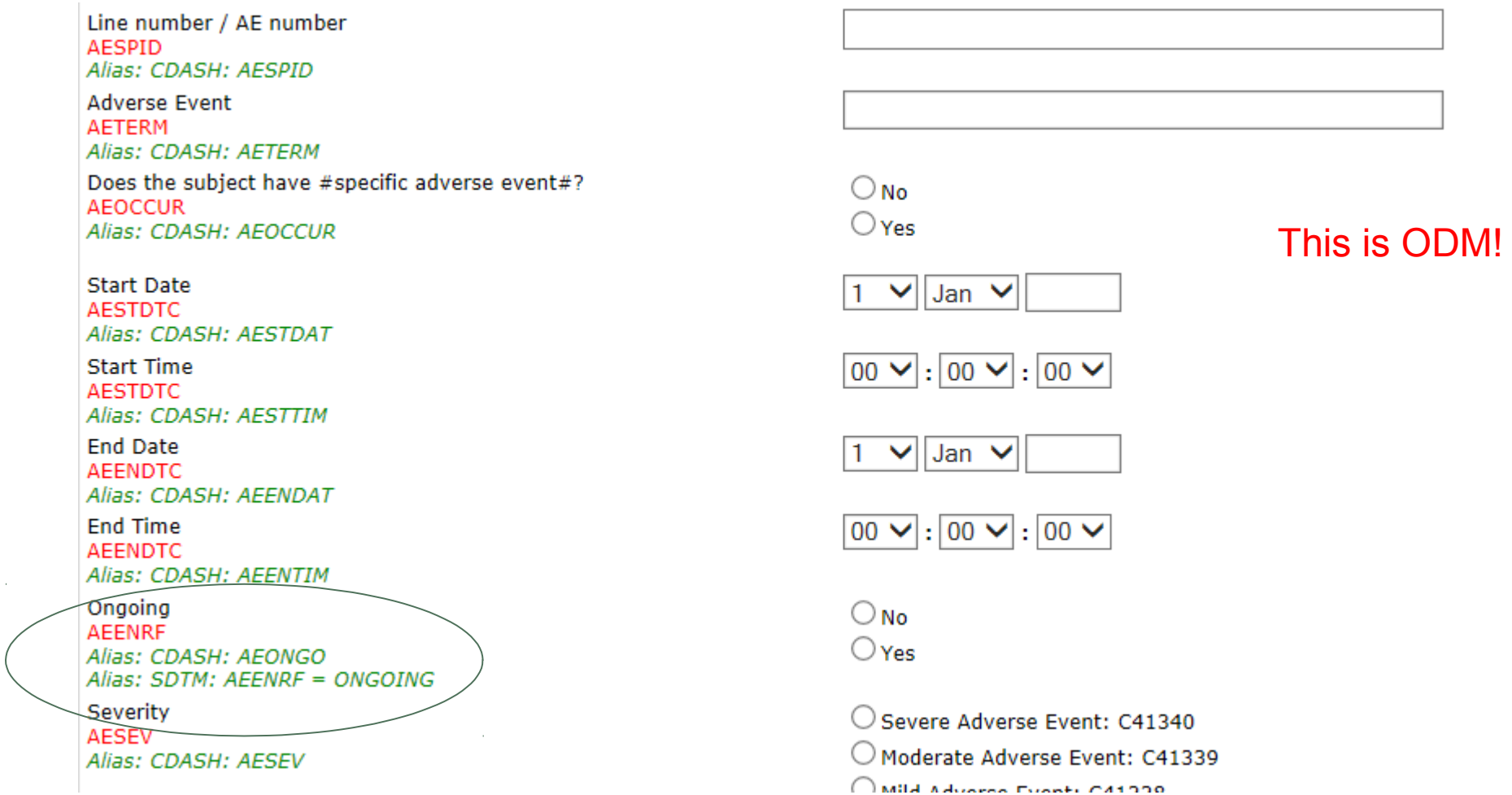
RESTful web service in define-XML

```
<ItemDef OID="IT.AE.AEACN" Name="AEACN" DataType="text" Length="30" SASFieldName="AEACN">
    <Description>
      <TranslatedText xml:lang="en">Action Taken with Study Treatment</TranslatedText>
    </Description>
    <!-- get the controlled terminology from SHARE -->
    <CodeListRef Reference="http://share.cdisc.org/rest/CDISC-CT/2015-06-26/CL.ACN"/>
    <!-- we used the CDASH form with Item "Action Taken with Study Treatment"
    so we just reference it using an HTTP GET request -->
    <def:Origin Type="CRF">
      <def:DocumentRef
        Reference="http://mycompany.com/rest/StudyRepository/CDASH/v.1.1/Form.AE/IG.AE_DETAILS/AE_22" />
    </def:Origin>
</ItemDef>
```

Consequences for ODM and for Define-XML

- (meta)data quality improvement
- Define-XML: annotated CRF references the SDTM-annotated ODM-XML Study Design, not a PDF
  - PDF is not really machine-readable
- Define-XML "Origin" can point to a SHARE CDASH data item
- With a simple tool, the reviewer can stil get a visual representation

Annotated CRF is the SDTM-annotated (ODM-XML) Study Design



With special thanks to David Iberson-Hurst, Assero

FDA: Files versus Databases & Services

- Large amounts of data belong in databases
  (files are just there for transport - and even then…)

- However, most reviewers are using files for review
  - Can you really inspect >106 rows in a table?
  - FDA should forbid reviewers to use files, and only allow "select" requests to the Janus-CTR warehouse
  - Reviewers should not be able to download files from Janus-CTR, only data

- Shouldn't we rethink the concept of "file exchange"?
  - We do not download files with book reviews from Amazon either…

Steps forward: the easy ones
(if we really want)

- Get rid of SAS-XPT now!
  - It is a silo - only used in Clinical Research

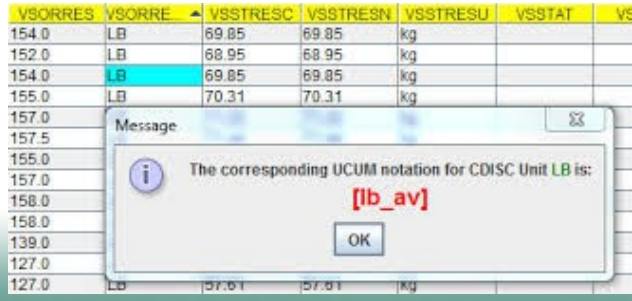
- Start using Dataset-XML and let it evolve
  - Can e.g. carry EHR data points and audit trail information

- Make SDTM more flexible
  - Less variables = more quality
  - more freedom for labels, …

- Allow LOINC and UCUM
  - And make them mandatory after 5 years

- Start working on ODM 2.0
  - Do learn from FHIR!

```
- <ItemData ItemOID="I_HEIGHT" Value="193">
    <MeasurementUnitRef MeasurementUnitOID="MU_CM" />
  - <cda:observation classCode="OBS" moodCode="EVN">
      <cda:templateId root="2.16.840.1.113883.10.20.1.31" />
      <!-- Result observation template  -->
      <cda:id root="d11275e1-67ae-11db-bd13-0800200c9a66" />
      <cda:code code="50373000" codeSystem="2.16.840.1.113883.6
        displayName="Body height" />
      <cda:statusCode code="completed" />
      <cda:effectiveTime value="20100313" />
      <cda:value xsi:type="PQ" value="193" unit="cm" />
    </cda:observation>
  </ItemData>
```



Steps forward: the harder ones …

- How much milliseconds does it take you to update your software when a new SDTM-IG is published?
  - Currently (estimated): 109-1011 ms
  - It should be ~102 ms

- Can we get a machine-readable IG please?
  - The SHARE template define-XML already helps a lot … (no more copy-and-paste…)

- Make SDTM more precise: how should machines interprete statements like "*The following Qualifiers would not generally be used in QS: ..., --METHOD, ...*"?

Steps forward: the harder ones …

- Many years ago, it was said that there would never be more than 20 SDTM domains …
- The content and order of variables in each SDTM domain is strictly regulated - WHY?
  - This is an insult to our intelligence …
- The "Guide" has become the "Book of Law" …
  - Sometimes feels like the Inquisition …
- Can't we do with less domains that are more flexible?

The train has left the station
But is it on the right track?

Steps forward: the harder ones … Moving away from tables

- The world is not flat (A. Oliva - FDA)
- But HL7-v3 wasn't the solution either …
  - The world should not become a Borg cube…
- We now have Dataset-XML
  - Which is again … flat
  - But we can easily make in multi-dimensional
- Can't we really do smarter?
  - Can we learn from FHIR?

A Borg cube …

"Smart" Dataset-XML for SDTM

Grouping by subject and visit

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Dataset-XML: Untyped Data Example  -->
<ODM
  xmlns="http://www.cdisc.org/ns/odm/v1.3"
  xmlns:xlink="http://www.w3.org/1999/xlink"
  xmlns:data="http://www.cdisc.org/ns/Dataset-XML/v1.0"
  FileType="Snapshot"
  ODMVersion="1.3.2"
  data:DatasetXMLVersion="1.0.0"
  FileOID="www.cdisc.org.Studycdisc01-Define-XML_2.0.0(IG.LB)"
  CreationDateTime="2016-01-19T09:31:03">
    <ClinicalData StudyOID="CDISC01" MetaDataVersionOID="MDV.CDISC01.SDTMIG.3.1.2.SDTM.1.2"
        Dataset="LB" Domain="LB">
        <!-- Dataset (LB) ordered by Subject and Visit -->
        <!-- We do not need STUDYID, DOMAIN, USUBJID, DOMAIN anymore as data point
            as we already grouped by them -->
        <SubjectData SubjectKey="CDISC01.100008">
            <StudyEventData StudyEventOID="SCREEN">
                <!-- LBSEQ is given by ItemGroupRepeatKey  -->
                <ItemGroupData ItemOID="IG.LB" ItemGroupRepeatKey="1">
                    <ItemData ItemOID="IT.LB.LBREFID" Value="B232115"/>
                    <ItemData ItemOID="IT.LB.LBTESTCD" Value="BILI"/>
                    <ItemData ItemOID="IT.LB.LBCAT" Value="CHEMISTRY"/>
                    <ItemData ItemOID="IT.LB.LBORRES" Value="0.4"/>
                    <ItemData ItemOID="IT.LB.LBORRESU" Value="mg/dL"/>
                    <ItemData ItemOID="IT.LB.LBORNRLO" Value=".0"/>
                    <ItemData ItemOID="IT.LB.LBORNRHI" Value="1.1"/>
                    <ItemData ItemOID="IT.LB.LBSTRESC" Value="6.8"/>
                    <ItemData ItemOID="IT.LB.LBSTRESN" Value="6.8"/>
```

Provisional…

Steps forward: EHR Integration

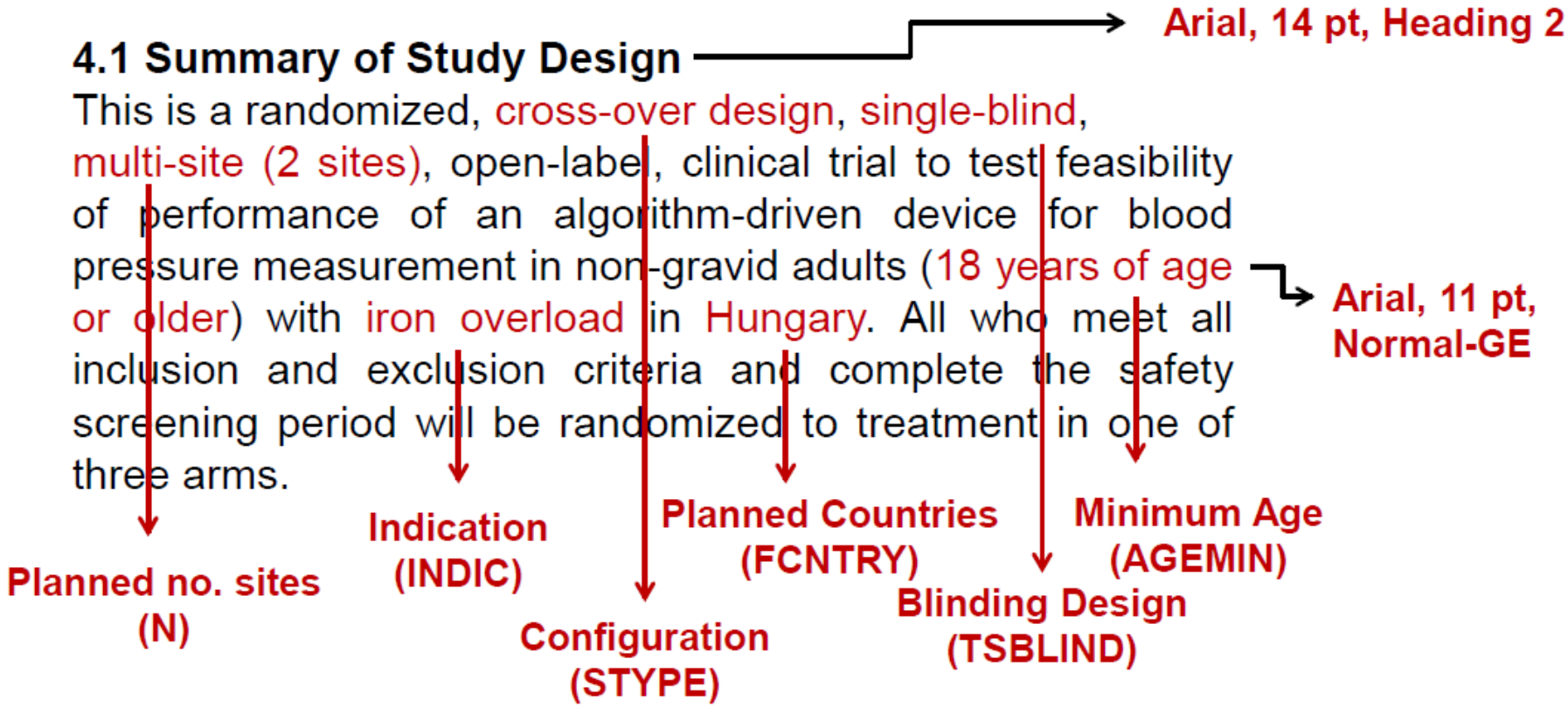
- We cannot do EHR integration unless we get rid of SAS Transport 5
- ODM / Dataset-XML can already carry EHR data points
- ODM-XML can ensure carrying EHR information from data capture to submission

```
- <ItemData ItemOID="I_HEIGHT" Value="193">
    <MeasurementUnitRef MeasurementUnitOID="MU_CM" />
  - <cda:observation classCode="OBS" moodCode="EVN">
      <cda:templateId root="2.16.840.1.113883.10.20.1.31" />
      <!-- Result observation template -->
      <cda:id root="d11275e1-67ae-11db-bd13-0800200c9a66" />
      <cda:code code="50373000" codeSystem="2.16.840.1.113883.6
        displayName="Body height" />
      <cda:statusCode code="completed" />
      <cda:effectiveTime value="20100313" />
      <cda:value xsi:type="PQ" value="193" unit="cm" />
    </cda:observation>
  </ItemData>
```

Steps forward: the hardest one … The machine-readable protocol

- Automates many steps in the process
- Allows to pick up forms (and other things) from an MDR automatically
  - These forms are of course SDTM-annotated
- Allows to set up the EDC system automatically (like the ODM-SDM-XML does now)
- But is still "protocol writer friendly"!!!

The hardest one …
The machine-readable protocol



Courtesy Angela Johnson, GE Healthcare, CDISC Chicago Interchange, 2015

What I did not talk about (yet)

- Using the cloud …
- Mobile health, mobile clinical research
  - FHIR + ODM "a marriage blessed in heaven"?
- Semantic Web and RDF
- NullFlavors and other monsters in SDTM
- No-SQL databases
  - E.g. native XML databases
- And many more things …

Links

- XML4Pharma's CDISC Web Services testbed (please try out in your applications)
  - ~ 20 CDISC-CT Services
  - 5 SDTM/SEND Domain-Variable services
  - For different versions of the standards
  - 3 LOINC services
  - 2 UCUM unit conversion services
  - CDISC and FDA rules as Xquery (experimental)
  - www.xml4pharmaserver.com
- National Library of Medicines Web Services (MedlinePlus Connect)
  - SNOMED-CT, ICD-9, ICD-10, RXCUI, NDC, LOINC
  - https://www.nlm.nih.gov/medlineplus/connect/overview.html

Links - Blogs

- Working on and with CDISC Standards
- CDISC end-to-end
- Reimagine Research - Thoughts on Improving Clinical Research
- Thoughts on Medical Informatics

One can share information between people using websites

One can SHARE information between applications using web services

But we need make our standards ready for this paradigm change

…

Slides at: www.XML4Pharma.com/slides/

Thank you for your attention

CDISC SHARE

SHARED HEALTH AND CLINICAL RESEARCH ELECTRONIC LIBRARY