

PINNACLE²¹

INDUSTRY METRICS FOR EXTENSIONS OF CDISC TERMINOLOGY

Sergiy Sirichenko



June 14th–18th

US Connect 2021

The Clinical Data
Science Conference

P²¹

SERGIY SIRICHENKO



DIRECTOR, CONSULTATIVE SERVICES

- ▶ Co-founder of OpenCDISC/Pinnacle 21
- ▶ SME on FDA JumpStart, coreDF and eDATA projects
- ▶ User advocate

P21

AGENDA

- Introduction
- Methodology
- Results
- Summary

P²¹



P21

INTRODUCTION

Why do we need industry metrics for extensions of CDISC Controlled Terminology?

CDISC CT SUMMARY

- ▶ Standard terms for standard structure
 - ▶ SDTM, SEND, ADaM, Define-XML
- ▶ Extensible and non-extensible codelists
- ▶ Valid extension
 - ▶ Not synonyms of std terms
 - ▶ No splitting or merging of std terms
 - ▶ Consistency within study and submission
 - ▶ Submitting request to CDISC for adding new terms
 - ▶ Correct type of information

TOP 15 VALIDATION ISSUES IN SDTM

Rule	Message	Type	Studies Failed	Studies with Unresolved	Issues Rate
CT2002	Variable value not found in extensible codelist	Warning	99%	99%	43.8%
SD1076	Model permissible variable added into standard domain	Warning	99%	99%	8.3%
SD1117	Duplicate records	Warning	86%	81%	16.7%
SD1078	Permissible variable with missing value for all records	Notice	82%	67%	37.8%
SD0029	Missing value for --STRESU, when --STRESC is provided	Warning	74%	68%	25.6%
SD0026	Missing value for --ORRESU, when --ORRES is provided	Warning	72%	67%	25.4%
DD0084	Referenced File is missing	Error	71%	23%	100.0%
SD0021	Missing End Time-Point value	Warning	69%	60%	21.8%
SD0037	Value for variable not found in user-defined codelist	Error	66%	21%	35.6%
SD0063	SDTM/dataset variable label mismatch	Warning	60%	36%	8.3%
SD0080	AE start date is after the latest Disposition date	Error	58%	43%	15.4%
SD0031	Missing values for --STDTC, --STRF and --STRTPT, when --ENDTC, --ENRF or --ENRTPT is provided	Warning	57%	50%	28.5%
CT2005	Variable value not found in extensible codelist when value-level condition occurs	Warning	57%	43%	43.8%
SD1231	Variable value is longer than defined max length when value-level condition occurs	Error	55%	25%	69.0%
CT2003	Coded and Decoded values do not have the same Code in CDISC CT	Error	55%	24%	12.9%

TOP CT2002 FAILURES BY VARIABLE

By studies with the issue

Variable	Dataset	Studies Failed	Issues Rate
LBTEST	LB	84%	7%
LBTESTCD	LB	84%	7%
LBORRESU	LB	73%	17%
PCORRESU	PC	69%	84%
LBSTRESU	LB	67%	14%
PCSTRESU	PC	66%	84%
CMDOSU	CM	59%	12%
TSPARM	TS	56%	8%
EPOCH	TA	55%	55%
EGTEST	EG	54%	43%
EGTESTCD	EG	54%	43%
CMDOSFRQ	CM	51%	13%
EPOCH	SE	50%	50%
EPOCH	LB	50%	50%
RACE	DM	49%	6%
EPOCH	AE	48%	72%
EPOCH	VS	48%	60%
EPOCH	DS	47%	33%
EPOCH	SV	42%	56%
EPOCH	CM	42%	44%
EPOCH	EX	40%	88%
QSCAT	QS	39%	79%

By issue rate

Variable	Dataset	Studies Failed	Issues Rate
DOMAIN	EC	18%	100%
SCTEST	SC	27%	97%
EPOCH	PP	15%	94%
PCSTRESU	PC	66%	84%
EXDOSFRM	EX	11%	80%
QSCAT	QS	39%	79%
EXLOC	EX	10%	72%
DATEST	DA	14%	67%
RPTEST	RP	13%	66%
RSTEST	RS	11%	65%
PCSPCCND	PC	13%	58%
RSSTRESC	RS	12%	56%
EGTEST	EG	54%	43%
TULOC	TU	11%	39%
SUDOSU	SU	23%	35%
VSSTRESU	VS	21%	30%
LBNRIND	LB	10%	29%
PPSTRESU	PP	21%	25%
LBSPEC	LB	16%	25%
PPTEST	PP	31%	25%
PRLOC	PR	10%	20%
CMDOSFRM	CM	11%	19%

COMPLAINS ABOUT P21 VALIDATION

- ▶ Why do we need to explain our CT extension if it's allowed?
 - ▶ by CDISC
 - ▶ already listed in Define.xml file
- ▶ CT2002 rule should be a notice, or not exist as a warning
 - ▶ P21 validation message Types
 - ▶ *Error*
some data problem with ~100% confidence
 - ▶ *Warning*
signaling about potential problem, manual diagnostics is expected to confirm
 - ▶ *Notice*
report to confirm correct implementation

P21

MAJOR QUESTION TO ANSWER

- ▶ How good is industry implementation of CDISC extensible CT?
 - ▶ Do we still need CT2002?
 - ▶ What are potential common issues and their prevalence?
 - ▶ Sources of problems?
 - ▶ How to improve validation?
 - ▶ Any current deficiency in standards?
 - ▶ What should good implementation practice focus on?



P21

METHODOLOGY

Design of industry metrics study

SCOPE OF RESEARCH

- Use of industry metrics
 - Test methodology
- Common problems in CT compliance
 - Select **some** issues for detailed review
 - Prevalence
 - Source
 - Expect analysis challenges due to non-standardized data
- Potential solutions

SELECTION CRITERIA

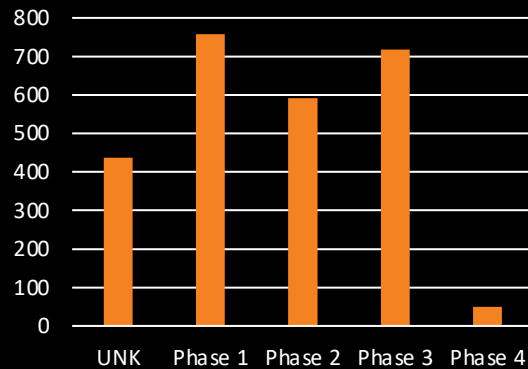
- P21 Enterprise metrics
- SDTM
- Last validation in 2020
- Number of datasets > 15
- Max 40 terms per variable
 - With some exceptions like --TEST, QNAM, ...
- Additional cleaning
 - ADaM datasets in SDTM validation, non-English data, ...

P21

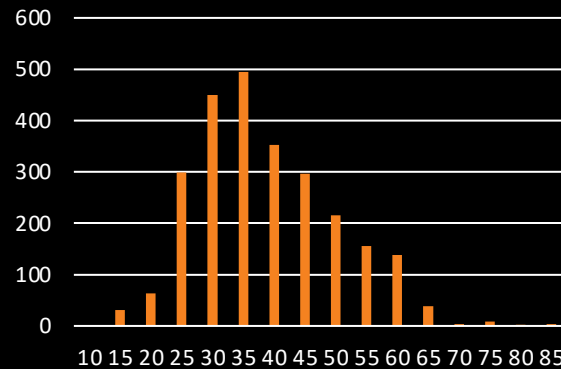
METRICS DATA

- ▶ 2,553 studies
- ▶ 106.5K datasets
- ▶ 492K codelists
- ▶ 3.55M terms

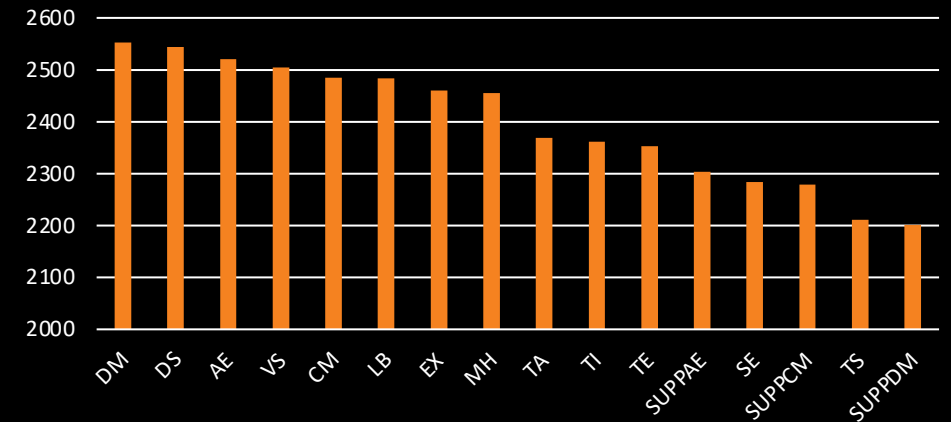
Studies by Phase



Studies with # of datasets



Frequency of datasets





RESULTS

RACE

- Non-extensible Codelist
- 58% of studies include additional terms
 - MULTIPLE – 28%
 - *MIXED, BIRACIAL*, etc.
 - OTHER – 29%
 - *OTHER ASIAN*, etc.
 - UNKNOWN – 14%
 - *NOT REPORTED, NOT PERMITTED, NOT APPLICABLE*, etc.
 - More granular Race – 0.7%
 - *JAPANESE, MAORI, INDIAN*, etc.
 - Synonyms of standard terms – 1.5%
 - *BLACK, Caucasian, Asian, Asia, 亚洲人*, etc.
 - Invalid info – 0.5%
 - *HISPANIC*

MULTIPLE

- RACE – 28%
- Combinational study drugs
 - AEACN – 16%
 - AEACN1, AEACN2 in SUPPAE
 - AEREL – 5.5%
- Alternative implementations
 - AEACN = '*Drug XYZ: DOSE REDUCED*' – 1.2%
 - RACE="" with all details in SUPPDM for cases of Other, Multiple or Unknown Race
- Solutions?
 - Add to CDISC CT
 - Create best practice
 - Adjust validation

OTHER

- ▶ Non-standard term
 - ▶ For info collected as free text
- ▶ What is CDISC/FDA guidance?
 - ▶ <free text> in domain
 - ▶ 'OTHER, <free text>'
 - ▶ 'OTHER', no details
 - ▶ 'OTHER' and details in SUPPQUAL
- ▶ Presence in variables
 - ▶ CMLOC – 48% studies
 - ▶ CMDOSFRQ – 53%
 - ▶ CMDOSFRM – 23%
 - ▶ CMDOSU – 45%
 - ▶ 'OTHER, <free text>' – 1%
 - ▶ Details in SUPPCM – 11%

AEACN

- Non-extensible Codelist
- 19% studies include additional terms
 - *MULTIPLE* (drugs) – 17%
 - Including “*Drug XYZ – DOSE REDUCED*”
 - Extended terms – 0.5%
 - *DOSE DELAYED*
 - *OTHER* – 1.2%
 - Synonyms of standard terms – 2%
 - *Dose not changed, NONE, NO ACTION TAKEN*
 - Invalid info – 0.7%
 - *DOSE MODIFIED, CONMED TAKEN, CONTINUING ON STUDY, REMEDIAL DRUG THERAPY, REQUIRED PROCEDURE, Y*

AEREL

- ▶ No standard CT
- ▶ Challenge with combinational drugs
- ▶ 100+ unique terms
 - ▶ Examples of rare terms:
 - ▶ *SUSPECTED, UNKNOWN, CERTAIN, NONE, NOT ASSESSABLE, REMOTE, Not Done, VERY RELATED, UNCLASSIFIABLE, PRIOR TO STUDY MEDICATION, etc.*

Term	Studies
NOT RELATED	55.8%
RELATED	45.6%
N	22.1%
Y	20.9%
POSSIBLY RELATED	13.1%
POSSIBLE	8.3%
PROBABLY RELATED	7.0%
UNRELATED	7.0%
UNLIKELY RELATED	6.9%
PROBABLE	6.8%
MULTIPLE	5.3%
UNLIKELY	4.7%
NO REASONABLE POSSIBILITY	3.8%
REASONABLE POSSIBILITY	3.8%
DEFINITE	2.6%
DEFINITELY RELATED	2.6%
NA	2.2%
DOUBTFUL	2.1%

VSTESTCD = VSALL

- ▶ When all assessments were not done for visit
- ▶ VSALL – 37% studies
 - ▶ 43 terms for VSTEST: *All Vital Signs, VS Data, All Tests, ...*
- ▶ LBALL – 35%
 - ▶ 52 terms for LBTEST
 - ▶ 55 another LBTESTCD terms with text 'ALL' (1.5% studies)
 - ▶ *ALLTESTS, CHEMALL, LBALLH, PCALL, ...*
- ▶ QS - 27%
 - ▶ + 3% Questionnaire-specific
 - ▶ *ECOGALL, HAMAALL, EQ5ALL, ...*

EPOCH

- ▶ 53% studies have non-standard terms for EPOCH
 - ▶ 2.8K unique terms
- ▶ Numbers
 - ▶ *BASELINE 4, BLINDED TREATMENT 3, CYCLE119, FOLLOW-UP 5, Part B - XYZ Dose Expansion for ABC Cycle 43, ...*
 - ▶ Cycles for oncology studies
 - ▶ Cross-over studies
- ▶ Invalid terms
 - ▶ *DRUG-FREE, EARLY TERMINATION, FOLLOW-UP PERIOD, INDUCTION, ...*

--ORRESU VS. --STRESU

- ▶ Original units are submitted as raw data instead of conversion to standard terms
- ▶ LB example

LBORRESU

Stats	Std. terms	Non-Std. terms	% of Std. terms
MEAN	12.7	4.0	77%
MEDIAN	12	3	79%
STDDEV	5.4	4.2	21%
MAX	31	28	100%
75%-tile	16	6	100%
25%-tile	9	0	64%

LBSTRESU

Stats	Std. terms	Non-Std. terms	% of Std. terms
MEAN	12.9	3.3	83%
MEDIAN	12	2	88%
STDDEV	5.9	5.6	19%
MAX	42	99	100%
75%-tile	15	4	100%
25%-tile	9	0	75%

METHODOLOGY PROBLEM

- ▶ Not all studies/datasets (N=2507) provided input to metrics due to cut-off limit for number of terms for variable

LBORRESU (N=1976)

Stats	Std. terms	Non-Std. terms	% of Std. terms
MEAN	12.7	4.0	77%
MEDIAN	12	3	79%
STDDEV	5.4	4.2	21%
MAX	31	28	100%
75%-tile	16	6	100%
25%-tile	9	0	64%

LBSTRESU (N=2417)

Stats	Std. terms	Non-Std. terms	% of Std. terms
MEAN	12.9	3.3	83%
MEDIAN	12	2	88%
STDDEV	5.9	5.6	19%
MAX	42	99	100%
75%-tile	15	4	100%
25%-tile	9	0	75%

- ▶ These 21% studies have hundreds of non-standard terms (*'as collected'*) for LBORRESU

EXAMPLE

PINNACLE²¹ ENTERPRISE

Unit Analysis

Search: lb gluc

Print Copy Download

Domain ^	TESTCD ^	TEST ^	ORRESU ^	STRESU ^	Count ^
LB	GLUC	Glucose	mg/dL	mmol/L	3,310
LB	GLUC	Glucose	MG/DL	mmol/L	2,199
LB	GLUC	Glucose	mmol/L	mmol/L	1,683
LB	GLUC	Glucose	MMOL/L	mmol/L	1,563
LB	GLUC	Glucose	mg/dl	mmol/L	754
LB	GLUC	Glucose	mmol/l	mmol/L	518
LB	GLUC	Glucose	g/L	mmol/L	365
LB	GLUC	Glucose	G/L	mmol/L	161
LB	GLUC	Glucose	g/l	mmol/L	135
LB	GLUC	Glucose	g/l	Scalar	21
LB	GLUC	Glucose	mg/dL	Scalar	17
LB	GLUC	Glucose	MG/DL	Scalar	7
LB	GLUC	Glucose	mmol/L	Scalar	3
LB	GLUC	Glucose	g/dl	Scalar	2

Found 14 records (filtered from a total of 524)

PINNACLE²¹ ENTERPRISE

Unit Analysis

Search: lb baso

Print Copy Download

Domain ^	TESTCD ^	TEST ^	ORRESU ^	STRESU ^	Count ^
LB	BASO	Basophils	10^9/L	10^9/L	3,093
LB	BASO	Basophils	10^3/uL	10^9/L	1,184
LB	BASO	Basophils	x10E9/L	10^9/L	685
LB	BASO	Basophils	10^3/UL	10^9/L	676
LB	BASO	Basophils	Giga/L	10^9/L	437
LB	BASO	Basophils	/mm^3	10^9/L	344
LB	BASO	Basophils	10^3/mm^3	10^9/L	323
LB	BASO	Basophils	K/uL	10^9/L	292
LB	BASO	Basophils	/mm^3	10^9/L	220
LB	BASO	Basophils	X10E9/L	10^9/L	211
LB	BASO	Basophils	GI/L	10^9/L	165
LB	BASO	Basophils	10^3/uL	10^9/L	153
LB	BASO	Basophils	thou/uL	10^9/L	149
LB	BASO	Basophils	x10E3/uL	10^9/L	107
LB	BASO	Basophils	10^9/L	10^9/L	66
LB	BASO	Basophils	/uL	10^9/L	50
LB	BASO	Basophils	cells/mm3	10^9/L	45
LB	BASO	Basophils	x10E3/mm3	10^9/L	42
LB	BASO	Basophils	thous/mm3	10^9/L	36
LB	BASO	Basophils	/ul	10^9/L	24
LB	BASO	Basophils	K/mm3	10^9/L	24
LB	BASO	Basophils	10**6/l	10^9/L	14
LB	BASO	Basophils	10**3/uL	10^9/L	7

Found 24 records (filtered from a total of 524)

EXAMPLE (CONTINUED)

TEST	LBORRESU	Expected Std. Units
Basophils	/mm*3	10^9/L
Basophils	/mm^3	10^9/L
Basophils	10**3/ul	10^9/L
Basophils	10*3/uL	10^9/L
Basophils	10*3/UL	10^9/L
Basophils	10*9/L	10^9/L
Basophils	10^3/uL	10^9/L
Basophils	10^9/L	10^9/L
Basophils	cells/mm3	10^9/L
Basophils	G/L	10^9/L
Basophils	Giga/L	10^9/L
Basophils	K/mm3	10^9/L
Basophils	K/ul	10^9/L
Basophils	thou/uL	10^9/L
Basophils	x10E3/uL	10^9/L
Basophils	x10E9/L	10^9/L
Basophils	X10E9/L	10^9/L
Basophils	/uL	10^6/L
Basophils	/ul	10^6/L
Basophils	10**6/l	10^6/L
Basophils	10*3/mm*3	10^12/L
Basophils	thous/mm3	10^12/L
Basophils	x10E3/mm3	10^12/L

- ▶ All collected LBORRESU values should be represented by standard terms in SDTM
- ▶ CT2002 should report these records as Errors (synonyms of std. terms)

VARIABLES WITH NEED FOR CT

▶ SUCAT

Category	# of studies	% of studies
NICOTINE	894	80%
ALCOHOL	586	53%
DRUGS	103	9%
CAFFEINE	93	8%
SPECIAL DIET	23	2%

▶ SUDOSU (36% studies)

- ▶ *SIGARS, GUM, CHEW, E-CIGARETTES*
- ▶ *GLASS, SHOT*
- ▶ *LINES, BUMP, OUNCES, JOINT, PINCH, TAN, TIN*
- ▶ There are many invalid synonyms of standard terms:
 - ▶ *PIPES* instead of *PIPE*
 - ▶ *DRINKS* or *Drink* instead of *DRINK*

SCTEST

- ▶ 97% studies with SC domain include extended terms for SCTEST
 - ▶ 179 (4%) std. terms vs. 4002 non-standard terms in 1003 studies
 - ▶ 1177 unique SCTEST
- ▶ Examples:
 - ▶ Cohort (26%), Stratification (30%), Randomization Code (0.4%), Female Reproductive Status (13%), Screening Number (14%,), Protocol Amendment N (18%), ...
- ▶ Invalid extensions:
 - ▶ *Age at Baseline*

QSTEST

- ▶ 210 Codelists for different Questionnaire (12K+ std. terms) based on QSCAT
 - ▶ 41K+ actual terms for QSTESTCD, 45K+ for QSTESTCD/QSTEST
 - ▶ If QSCAT is not standard term, then validation is skipped
- ▶ Only 29% QSCAT values are std. CT
 - ▶ ECOG – 423 studies
 - ▶ All terms which include 'ECOG' text – 622
 - ECOG PERFORMANCE STATUS*
 - ECOGPS*
 - ECOG_V1982*
 - EASTERN COOPERATIVE ONCOLOGY GROUP PERFORMANCE STATUS (ECOG)*
 - ▶ Other terms:
 - EASTERN COOPERATIVE ONCOLOGY GROUP*
 - EASTERN COOPERATIVE ONCOLOGY GROUP PERFORMANCE STATUS SCALE*

CT2002 ISSUES IN VS DOMAIN

- ▶ 42% studies have CT extensions for **VSTEST**
 - ▶ Only **10%** when ignoring VSTESTCD=VSALL
 - ▶ **~60% reported validation issues are incorrect CT extension:**
 - ▶ *Age at Visit, Area, Baseline Weight, Blood Pressure, Body Circumference, 'Body Surface Area, As Collected', Body Temperature, Derived Height, ECOG Performance Status, Oral Temperature, Other test*
- ▶ 14% studies have CT extensions for VSSTRESU
 - ▶ 51 unique terms
 - ▶ Most extensions are invalid
 - ▶ Some terms to be added to CDISC CT VSRESU
 - ▶ *kPA* – need for PMDA submissions
 - ▶ No terms for volume. E.g., *L/min*, *L*

VSSTRESU	# of studies
BEATS/MIN	158
kPa	63
BREATHS/MIN	47
/min	18
CM	18
Beats per min	13
BPM	13
c	12
KG	12
L/min	12

CT2003 ERROR – CODE/DECODE

- ▶ Same NCI Code for *Code and *Name variables
- ▶ Common invalid explanation: *'extendable CT'*
- ▶ CT2003 issue is always mapping or programming error
- ▶ VSTESTCD=*BMI*
 - ▶ *Body Mass Index* – correct (98%)
 - ▶ *BMI, bmi, BODY MASS INDEX* – synonyms
 - ▶ *BMI CALCULATED IN HOUSE* – additional qualifiers
 - ▶ *BMI Z-Score* – different tests
 - ▶ *Weight* – different info
- ▶ Cannot automate validation for inconsistent info in case of non-standard terms

CT2003 ERROR FOR BOTH STD. TERMS

1. Inconsistent info
2. Inconsistent CDISC standard across versions

LBTESTCD	NCI code	LBTEST	LBTEST NCI code	LBTEST according to NCI code for LBTESTCD
BASOBLE	C130155	Basophils Band Form	C130154	Basophils Band Form/Leukocytes
CASTS	C74763	Hyaline Casts	C74770	Casts
CCPAB	C96595	Cyclic Citrullinated Peptide IgG Ab	C147316	Cyclic Citrullinated Peptide Antibody
CRYSTALS	C74673	Calcium Oxalate Crystals	C74670	Crystals
CXCL10	C112238	Chemokine (C-X-C Motif) Ligand 1	C128952	Chemokine (C-X-C Motif) Ligand 10
EGFR	C112273	Epidermal Growth Factor	C82009	Epidermal Growth Factor Receptor
EOSBLE	C114217	Eosinophils/Leukocytes	C64604	Eosinophils Band Form/Leukocytes
HS1IGGAB	C96666	Herpes Simplex Virus 1 IgG Antibody	C96697	Herpes Simplex Virus 1/2 IgM Antibody
LYMCE	C98751	Eosinophils/Total Cells	C98720	Lymphocytes/Total Cells
NACREAT	C79464	Sodium	C64809	Sodium/Creatinine
PH	C45997	Specific Gravity	C64832	pH
RBCNUCLE	C82046	Nucleated Erythrocytes/Erythrocytes	C74647	Nucleated Erythrocytes/Leukocytes
RH	C92948	Rheumatoid Factor	C74717	Rh Factor
UNSPCECE	C114225	Unspecified Cells	C112241	Unspecified Cells/Total Cells
UREAN	C125949	Blood Urea Nitrogen	C61019	Urea Nitrogen
YEASTBUD	C106504	Yeast Cells	C74664	Yeast Budding



P²¹

SUMMARY

CT METRICS

- ▶ Can help to
 - ▶ Identify
 - ▶ Diagnose
 - ▶ Measure industry implementation issues
- ▶ Should be carefully
 - ▶ Designed
 - ▶ Interpreted
- ▶ Cleaning is important
- ▶ Limitations for metrics automation
 - ▶ Unexpected violations of standard
 - ▶ Ongoing non-finalized projects
 - ▶ Free-text data



CT ISSUES AND THEIR RESOLUTIONS

- ▶ Invalid extension of CDISC CT is still a common case
- ▶ Education and promotion of good practice should help
- ▶ Further enhancement of CDISC CT is expected

CT VALIDATION

- ▶ Still needed to help with correct implementation of CT
- ▶ Further enhancements are expected
 - ▶ better diagnostics
 - ▶ more accurate results
 - ▶ new functionality

P²¹



THANK YOU ;)

Sergiy Sirichenko

sergiy@pinnacle21.com

P21