Good Data Validation Practices

PhUSE Boston SDE

Max Kanevsky
April 27, 2017

PINNACLE 21

Why do we need Good Data Validation Practices?

Example

- A large Sponsor submitted data to PMDA, but it was returned due to an issue that Sponsor was not aware of
- Triggered rejection rule CT2001 on BLFL
- How could this happen?

Example (cont.)

- BLFL value should be “Y” or null, but is assigned a NY codelist (N, NA, U, Y)
- P21 fixes this inconsistency by adding a Y codelist to P21 provided CT files
- Sponsor used custom CT files instead

Most submissions have problems with data validation

- Incorrect installation, configuration, or customization of Validator
- Using older versions of Validator
- Incorrect interpretation of validation results
- Invalid or irrelevant explanations for validation findings
- Not validating data vs. define.xml

Key reason: Lack of knowledge

- Simplicity of Validator may be misleading in term of its correct use
- There is a lack of documentation and education resources for data validation

Good Data Validation Practices

- What is data validation?
- How to configure Validator?
- How to perform data validation?
- How to interpret validation results?
- How to evaluate risk of data issues?
- How to fix data errors?
- How to explain data issues?

What is Data Validation?

FDA definition of Data Validation

- "Study data validation helps to ensure that the study data are
- compliant,
- useful, and will
- support meaningful review and analysis"

Data Validation and Standards

- Validator was created to help with implementation of CDISC standards
- Standardized data allows automation including data validation

Data Validation and Data Quality

- Changing landscape – focus is moving from Conformance to Data Quality
- Adoption of data standards is accomplished
- Standard Conformance is a pre-requisite for regulatory submission

Data Quality is What Matters

- Value of data is its content
- Data quality is defined by intended use
- All users are different

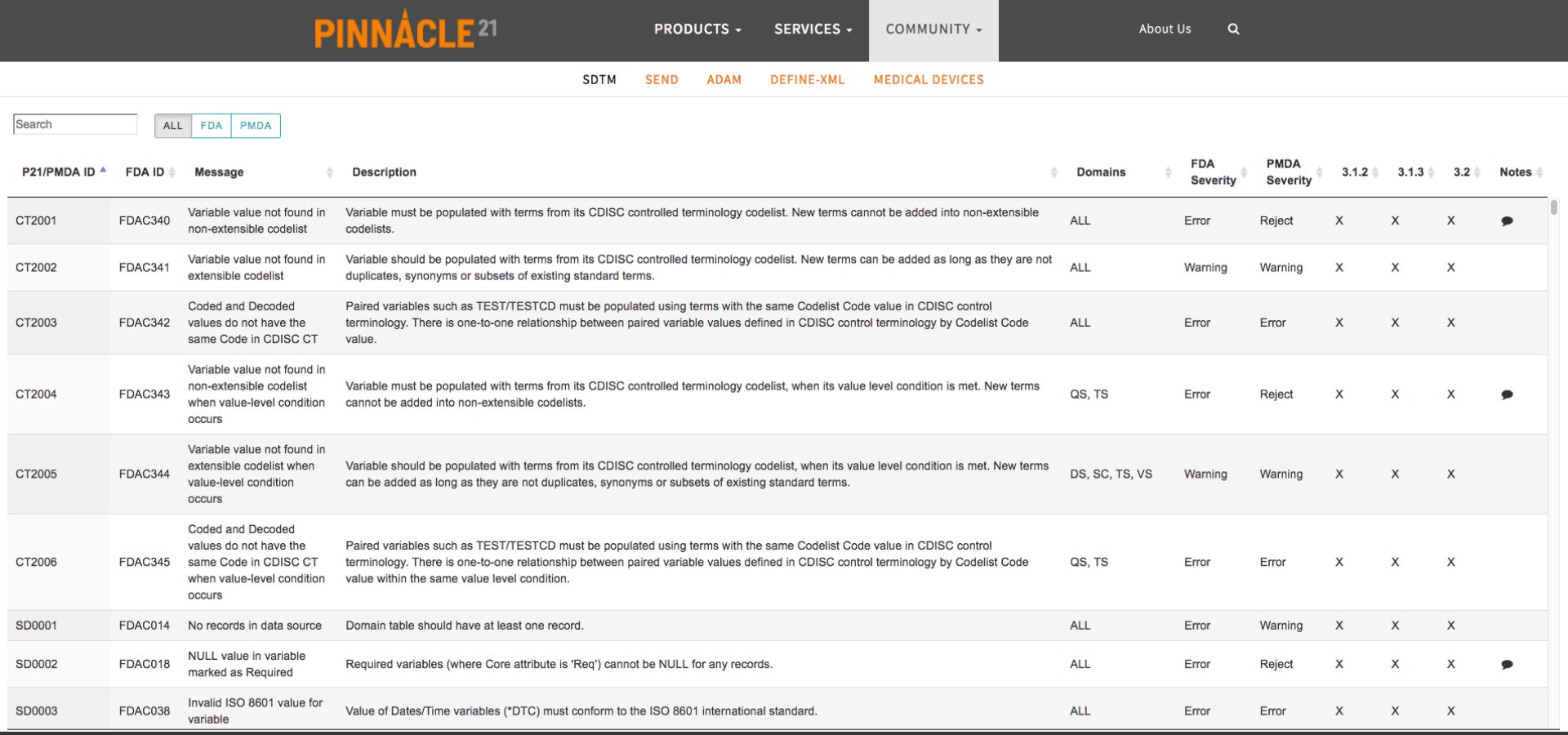
Data Quality by Intended Use

- FDA Clinical Trial Repository (Janus)
  - Can I load data?
- FDA JumpStart analyst
  - Can I run tools?
- Reviewer
  - Can data support meaningful review and analysis?

What is the source of validation rules?

Source of Validation Rules

- CDISC
  - Standards conformance rules
- FDA/PMDA
  - Technical Rejection Criteria
  - Business rules
- Sponsors
  - Study specific data quality rules



P21/PMDA ID	FDA ID	Message	Description	Domains	FDA Severity	PMDA Severity	3.1.2	3.1.3	3.2	Notes
CT2001	FDAC340	Variable value not found in non-extensible codelist	Variable must be populated with terms from its CDISC controlled terminology codelist. New terms cannot be added into non-extensible codelists.	ALL	Error	Reject	X	X	X	
CT2002	FDAC341	Variable value not found in extensible codelist	Variable should be populated with terms from its CDISC controlled terminology codelist. New terms can be added as long as they are not duplicates, synonyms or subsets of existing standard terms.	ALL	Warning	Warning	X	X	X	
CT2003	FDAC342	Coded and Decoded values do not have the same Code in CDISC CT	Paired variables such as TEST/TESTCD must be populated using terms with the same Codelist Code value in CDISC control terminology. There is one-to-one relationship between paired variable values defined in CDISC control terminology by Codelist Code value.	ALL	Error	Error	X	X	X	
CT2004	FDAC343	Variable value not found in non-extensible codelist when value-level condition occurs	Variable must be populated with terms from its CDISC controlled terminology codelist, when its value level condition is met. New terms cannot be added into non-extensible codelists.	QS, TS	Error	Reject	X	X	X	
CT2005	FDAC344	Variable value not found in extensible codelist when value-level condition occurs	Variable should be populated with terms from its CDISC controlled terminology codelist, when its value level condition is met. New terms can be added as long as they are not duplicates, synonyms or subsets of existing standard terms.	DS, SC, TS, VS	Warning	Warning	X	X	X	
CT2006	FDAC345	Coded and Decoded values do not have the same Code in CDISC CT when value-level condition occurs	Paired variables such as TEST/TESTCD must be populated using terms with the same Codelist Code value in CDISC control terminology. There is one-to-one relationship between paired variable values defined in CDISC control terminology by Codelist Code value within the same value level condition.	QS, TS	Error	Error	X	X	X	
SD0001	FDAC014	No records in data source	Domain table should have at least one record.	ALL	Error	Warning	X	X	X	
SD0002	FDAC018	NULL value in variable marked as Required	Required variables (where Core attribute is 'Req') cannot be NULL for any records.	ALL	Error	Reject	X	X	X	
SD0003	FDAC038	Invalid ISO 8601 value for variable	Value of Dates/Time variables (*DTC) must conform to the ISO 8601 international standard.	ALL	Error	Error	X	X	X	

Validation Rules Browser

Search across all available validation rules, locate rule-related information, and see what's required by FDA and PMDA

https://www.pinnacle21.com/validation-rules

Notice

Warning

Error

Reject

How to interpret Severity?

Pinnacle 21 Interpretation

- Errors are issues reported with 100% confidence
  - Start Date is after End Date
  - Missing SDTM required variable
- Warnings are potential issues requiring manual review
  - Terms added to Extensible CT
  - Missing Units on Results

P21

FDA/PMDA Interpretation

- Error/Warning – issues with impact on review process
- Reject - critical issues that prevent review

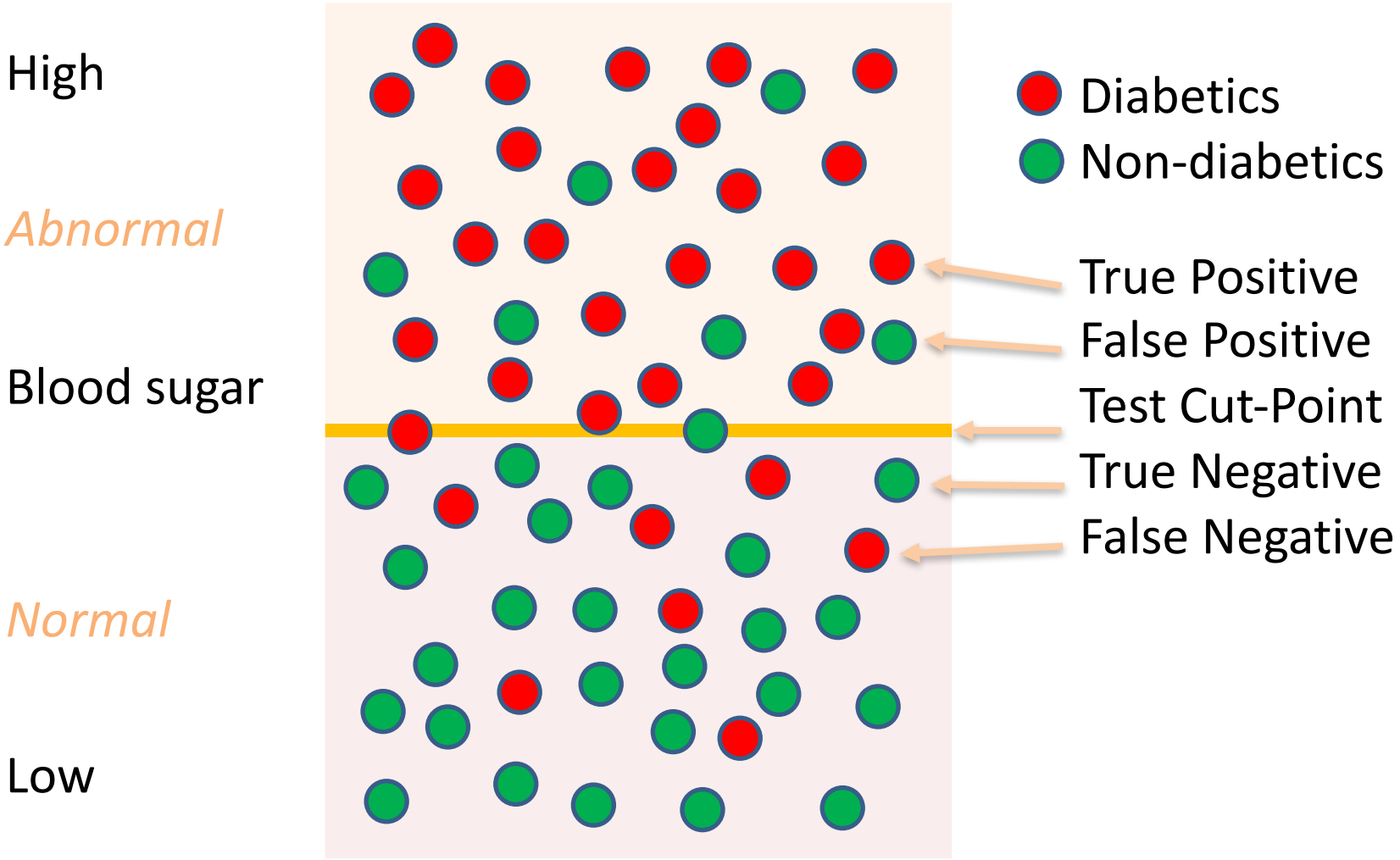
- Since Mar 14th, FDA no longer publishes Severity, except Rejection criteria

Why do we have False-Positive and False-Negative issues?

Issue exists, and is reported	Issue does not exist, but it is reported
Issue exists, but it is not reported	Issue does not exist, and it is not reported

True Positive	False Positive
False Negative	True Negative

Warnings produce both false-positive and potentially false-negative issues



Example: Blood sugar diagnostic test

P21

Balancing between FP and FN

- Rule tuning can change detection sensitivity of real data issues
- However, a decrease of false-positive issues may also increase the number of non-reported actual issues

P21

What to do?

- Minimizing false-positives when manual review is expensive
  - for example, extra work for data vendors or when missing planned timelines
- Minimizing false-negatives when impact of non-detected issue is high
  - for example, non-detected issue could delay regulatory review

P21

A few notes

- Almost nobody complains about false-negatives
- Programming bug is a bug, so report it to P21 team for fast fix
- Invalid business rule will result in invalid P21 rule, so report it to rule publisher

Get updates about GDVP

 linkedin.com/company/pinnacle-21

 twitter.com/pinnacle_21

 facebook.com/pinnacle21

 pinnacle21.com/blog

Summary & Questions

- Data validation is an important process
- Most submissions still have problems with data validation
- Lets work together toward “Good Data Validation Practices”

Max Kanevsky
Founder and CEO of Pinnacle 21
mkanevsky@pinnacle21.net