

A Learned Approach to CDISC Specifications

Brandon Welch
Stephen M. Noga
Jeff Abolafia

Introduction

- Why bother?
- Old School
- New School
- Experiment
- Results

Why bother?

- Spec'ing isn't fun – to us
- Why start from scratch when we have data?



Domain	varName	sortOrder	Label	Type	Length	OriginType	ProgDef
AE	STUDYID	1	Study Identifier	Char	9	Protocol	XXXXXXXX'
AE	DOMAIN	2	Domain Abbreviation	Char	2	Assigned	='AE'
AE	USUBJID	3	Unique Subject Identifier	Char	20	Derived	STUDYID - DM.SUBJID, EXAMPLE: 'XXXXXXXX-0008-00009'
AE	AESEQ	4	Sequence Number	Num	8	Derived	Sort dataset by key fields. For each record per subject, assign a sequential number from 1 to the number of records for that subject
AE	AEGRPID	5	Group ID	Char	40	CRF	



Old school



Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	AE	Identifier	Two-character abbreviation Req for the domain.	
USUBJID	Unique Subject Identifier	Char		Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Req

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	AE	Identifier	Two-character abbreviation Req for the domain.	
USUBJID	Unique Subject Identifier	Char		Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Req

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	AE	Identifier	Two-character abbreviation Req for the domain.	
USUBJID	Unique Subject Identifier	Char		Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Req

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	AE	Identifier	Two-character abbreviation Req for the domain.	
					Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Req

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	AE	Identifier	Two-character abbreviation Req for the domain.	
					Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Req

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	AE	Identifier	Two-character abbreviation Req for the domain.	
USUBJID	Unique Subject Identifier	Char		Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Req

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	AE	Identifier	Two-character abbreviation Req for the domain.	
USUBJID	Unique Subject Identifier	Char		Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Req

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	AE	Identifier	Two-character abbreviation Req for the domain.	
USUBJID	Unique Subject Identifier	Char		Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Req

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	AE	Identifier	Two-character abbreviation Req for the domain.	
USUBJID	Unique Subject Identifier	Char		Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Req

Old school

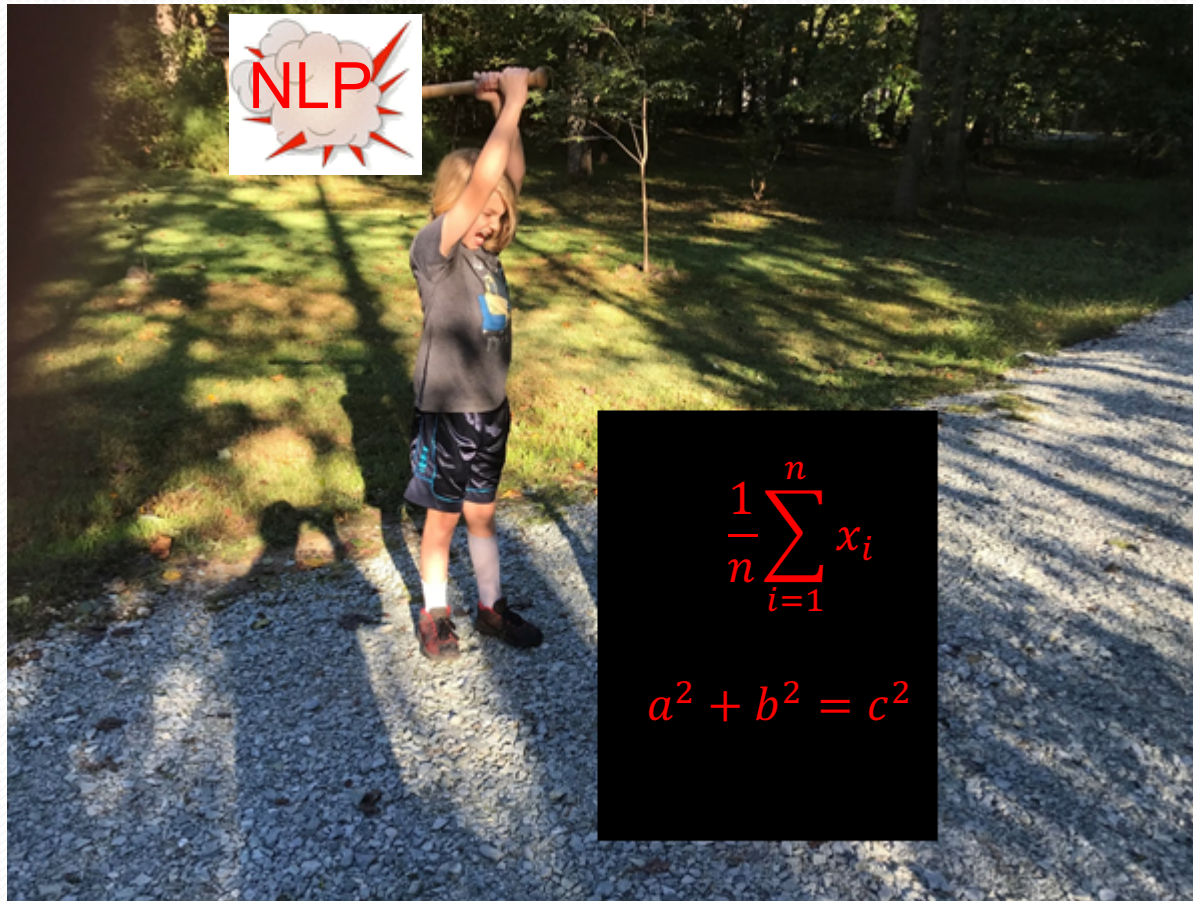


- Brute force/visual approach
- What SAS programmers have always used
- Using SAS
 - Frequencies
 - IF THEN/ELSE
 - Nothing fancy

TRTEMFL

	PROGDEF	 COUNT
	"Y" if ADAE.ASTDT >= ADSL.TRTSDT and ADSL.TRTSDT > .z missing otherwise	9
	"Y" if ADSL.TRTSDT <= ADAE.ASTDT =missing otherwise	7
	"Y" if ADAE.ASTDT >= ADSL.TRTSDT and ADSL.TRTSDT > .z, else missing	3
	"Y" if ASTDT >= TRTSDT and TRTSDT > .z else, missing.	2
	"Y" if .z < ADSL.TRTSDT <= ADAE.ASTDT <= ADSL.LSTVISDT =missing otherwise	2
	"Y" if >< ADSL.TRTSDT <= ADAE.ASTDT or ADAE.ASTDT is missing	2

New school – to us



New school – to us

- Text pre-processing

```
= 'Y' if ADSL.TRTSDT <= ASTDT <= ADSL.TRTEDT + 30 if  
      ADAE.RELGR1N = 0  
= 'Y' if ADSL.TRTSDT <= ASTDT if ADAE.RELGR1N = 1
```

REGEX



Python INFLECT module

```
if trtsdt less than or equal to astdt less than or equal to  
trtedt plus thirty if relgr1n equal to zero
```

```
equal to yes if trtsdt less than or equal to astdt if relgr1n  
equal to one
```

New school – to us

- Cosine similarity

Abbrev. Def.	Similarity value1	Similarity value2	Similarity value3	Record (1 st comparis on)	Record (2 nd comparis on)	Record (3 rd compari son)
if astdt greater than or equal to	0.950081	0.865946	0.834122	28	50	45
if astdt greater than or	0.950081	0.856613	0.854315	50	28	56
...						

New school – to us

- Sentence scoring

Definition 1:

if aedoseyn equal to yes then set to yes otherwise set to missing



Sum of weights



if	aedoseyn	equal	to	yes	then	set	to	yes	otherwise	set	to	missing	Score
0	0.33	0.67	0	1	0	0.67	0	1	0.67	0.67	0	0.33	5.33

Experiment

Using 3 recent ADaM studies, compared old school and new school:

Across all metadata

TRTEMFL

PROGDEF	COUNT	
"Y" if ADAE.ASTDT >= ADSL.TRTSDT and ADSL.TRTSDT > .z missing otherwise	9	1 slot
"Y" if ADSL.TRTSDT <= ADAE.ASTDT =missing otherwise	7	2 slot
"Y" if ADAE.ASTDT >= ADSL.TRTSDT and ADSL.TRTSDT > .z, else missing	3	3+ slot
"Y" if ASTDT >= TRTSDT and TRTSDT > .z else, missing.	2	
"Y" if .z < ADSL.TRTSDT <= ADAE.ASTDT <= ADSL.LSTVISDT =missing otherwise	2	
"Y" if > < ADSL.TRTSDT <= ADAE.ASTDT or ADAE.ASTDT is missing	2	

Experiment

Old School approach

TRTEMFL

PROGDEF	COUNT
"Y" if ADAE.ASTDT >= ADSL.TRTSDT and ADSL.TRTSDT > .z missing otherwise	9
"Y" if ADSL.TRTSDT <= ADAE.ASTDT =missing otherwise	7
"Y" if ADAE.ASTDT >= ADSL.TRTSDT and ADSL.TRTSDT > .z, else missing	3
"Y" if ASTDT >= TRTSDT and TRTSDT > .z else, missing.	2
"Y" if .z < ADSL.TRTSDT <= ADAE.ASTDT <= ADSL.LSTVISDT =missing otherwise	2
"Y" if .z < ADSL.TRTSDT <= ADAE.ASTDT or ADAE.ASTDT is missing	2

Study 1 def

Study 2 def

Study 3 def

Experiment

- Cosine

Abbrev. Def.	Sim. value1	Sim. value2	Sim. value3	Record (1 st comp)	Record (2 nd comp)	Record (3 rd comp)
if asdt greater than or equal to	0.950081	0.865946	0.834122	28	50	45
if asdt greater than or	0.950081	0.856613	0.854315	50	28	56
...						

Study 1 def

Study 2 def

Study 3 def

Experiment

- Scoring

Definition	Score
if trtsdt less than or equal to aesdt less than or equal to trtedt or trtsdt less than or equal to aesdt and trtedt is missing missing otherwise	12.7511
if asdt greater than or equal to trtsdt and trtsdt greater than missing equal to yes if asdt is missing equal to missing otherwise	11.0034
for vaccinees equal to yes if asdt greater than or equal to trtsdt equal to yes if asdt is missing for hhcs equal to missing	9.6728
if trtsdt less than or equal to aesdt or aesdt not missing equal to no otherwise equal to missing if anyae equal to no	9.6687
if trtsdt less than aesdt less than or equal to spaedt or aeonset equal to onset after first dose of study drug missing otherwise	9.2586

Study 1 def

Study 2 def

Study 3 def

Experiment

Analysis Dataset	Spec Position	# of Variables	Percent
ADAE	1	41	45.6
	2	28	31.1
	3+	21	23.3
ADCM	1	17	48.6
	2	10	28.6
	3+	8	22.9
ADDV	1	16	64.0
	2	8	32.0
	3+	1	4.0
ADEF	1	5	22.7
	2	5	22.7
	3+	12	54.5
ADEG	1	22	64.7
	2	1	2.9
	3+	11	32.4
ADEX	1	1	16.7
	2	1	16.7
	3+	4	66.7
ADLB	1	26	61.9

WAN Percentage!!!

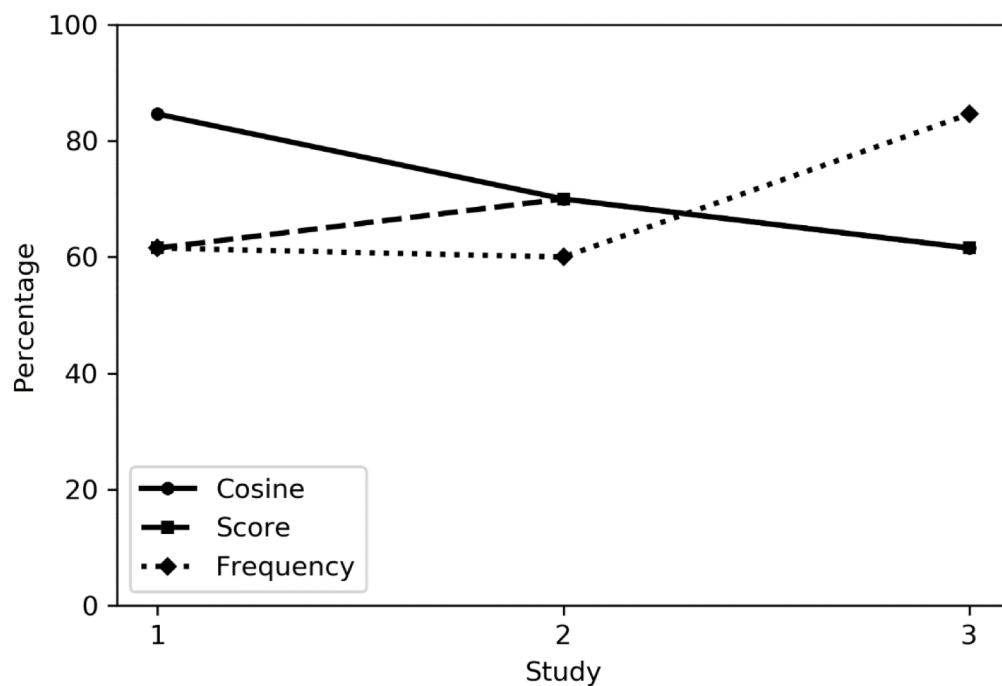
Welch
Abolafia
Noga

An illustration:

$$3/7 = 43\%$$

Results

Highest predicted definition



TRTEMFL

PROGDEF	COUNT
"Y" if ADAE.ASTDT >= ADSL.TRTSDT and ADSL.TRTSDT > .z missing otherwise	9
"Y" if ADSL.TRTSDT <= ADAE.ASTDT =missing otherwise	7
"Y" if ADAE.ASTDT >= ADSL.TRTSDT and ADSL.TRTSDT > .z, else missing	3
"Y" if ASTDT >= TRTSDT and TRTSDT > .z else, missing.	2
"Y" if .z < ADSL.TRTSDT <= ADAE.ASTDT <= ADSL.LSTVISDT =missing otherwise	2
"Y" if >< ADSL.TRTSDT <= ADAE.ASTDT or ADAE.ASTDT is missing	2

Conclusion

- Best of both worlds



The Big Problem With Machine Learning Algorithms

- “Machine learning algorithms will always identify a pattern, even if there is none”

Questions?

- Brandon Welch
- Rho, Inc.
- Email: bwelch@rhoworld.com
- Phone: 919-595-6592