

AI-powered unstructured data analysis to accelerate Biometrics activities

Hao Chen
Yiwen Wang
Kevin Huang



Agenda

Background

Business Problem

New Perspective

Case Sharing

Summary



Background

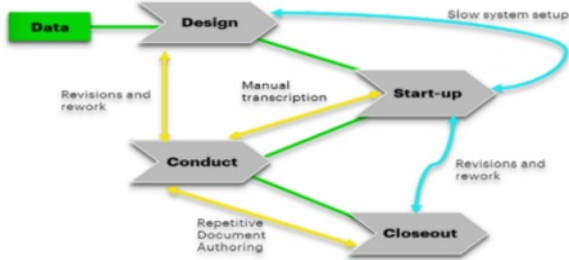


TRANSITIONING TO FUTURE OF CLINICAL TRIALS

Study published by the [Massachusetts Institute of Technology \(MIT\)](#) has found that only 13.8% of drugs successfully pass clinical trials company can expect to pay between \$161 million to \$2 billion for any drug to complete the entire clinical trials. This Artificial Intelligence has the potential to Significantly reduce clinical cycle time and cost while helping to improve the outcome of clinical development.

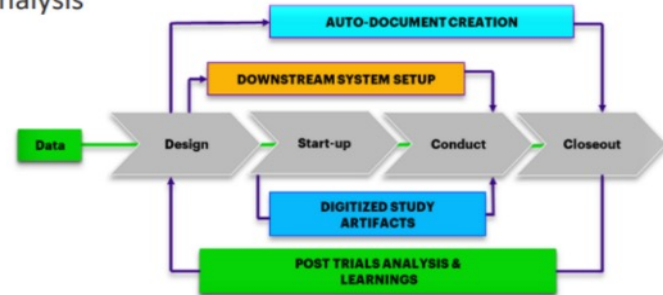
Traditional clinical trail data flow

A Non-linear maze of manual effort, rework
Inefficiency



The future state vision

Streamline interoperability of trial artifacts enabled through the standardization of data/vocabulary learnings from AI & ML analysis



FUTURE STATE BENEFITS

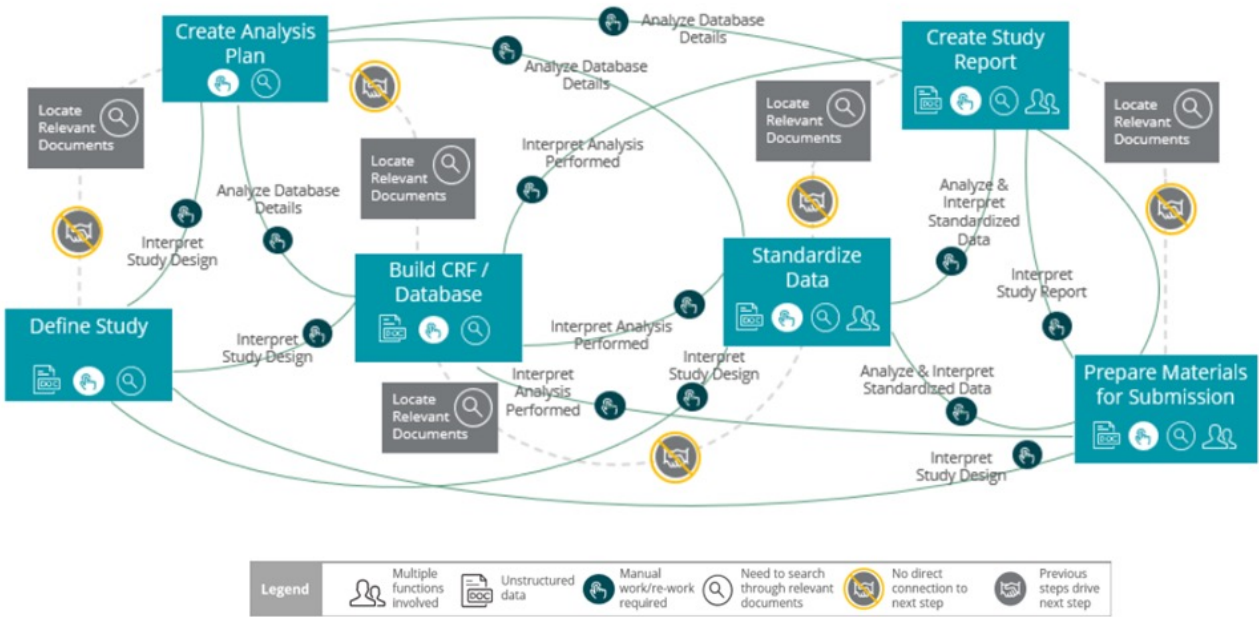
- Reduced manual Effort & rework.**
- Increased interoperability Across Multiple systems**
- Improved Trial Cycle time.**
- Optimized Use of Insights To Improve the Next Trial**
- Reduced Errors in Trial artifacts.**



Business Problem



Clinical Data Flow : Current State



Statistician



Programming



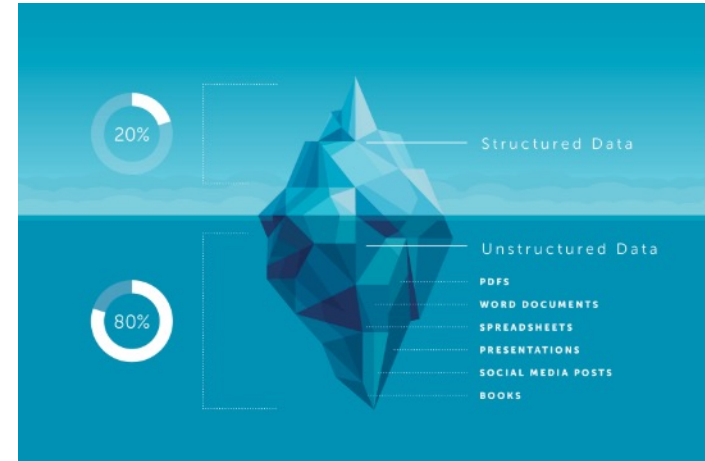
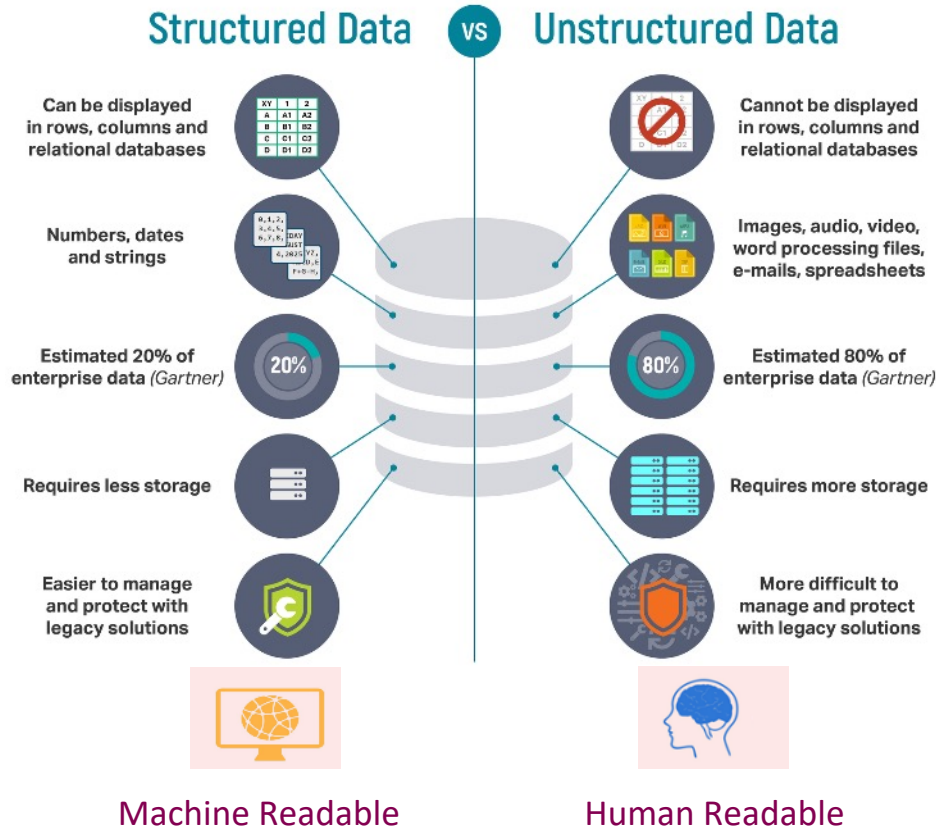
Data Management



Medical Communication



Structured Data vs Unstructured Data



Problem examples

1) Output Table analysis

File Format/Table Layout varies across different TAs/Vendors, how to extract the data?

Table 3. Patient Screening and Enrollment, Trials A and B

Disposition	Trial A	Trial B
Patients screened	XX	XX
Screening failures	XXX (XX.X%)	XXX (XX.X%)
Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
Other	XXX (XX.X%)	XXX (XX.X%)
Patients enrolled	XXX (XX.X%)	XXX (XX.X%)
Patients randomized	XXX (XX.X%)	XXX (XX.X%)



(.rtf/.doc/.docx)

Table 3. Patient Screening and Enrollment, Trials A and B

Disposition	Trial A	Trial B
Patients screened	XX	XX
Screening failures	XXX (XX.X%)	XXX (XX.X%)
Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
Other	XXX (XX.X%)	XXX (XX.X%)
Patients enrolled	XXX (XX.X%)	XXX (XX.X%)
Patients randomized	XXX (XX.X%)	XXX (XX.X%)



Table 3. Patient Screening and Enrollment, Trial

Disposition	Trial A	Trial B
Patients screened	XX	XX
Screening failures	XXX (XX.X%)	XXX (XX.X%)
Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
Other	XXX (XX.X%)	XXX (XX.X%)
Patients enrolled	XXX (XX.X%)	XXX (XX.X%)
Patients randomized	XXX (XX.X%)	XXX (XX.X%)



	A	B	C	D	E
1	rowNum	Lev0	Lev1	V2	V3
2		1 Disposition		Trial A	Trial B
3		2 Patients screened		XX	XX
4		3 Screening failures		XXX (XX.X%)	XXX (XX.X%)
5		4 Screening failures	Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
6		5 Screening failures	Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
7		6 Screening failures	Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
8		7 Screening failures	Other	XXX (XX.X%)	XXX (XX.X%)
9		8 Patients enrolled		XXX (XX.X%)	XXX (XX.X%)
10		9 Patients randomized		XXX (XX.X%)	XXX (XX.X%)



1) Output Table analysis

Table 1.1.2.1 Demographic Characteristics of Survey Respondents		PDF	Age (years)	Age (years)
			12.4	12.4
			Mean (SD) 67.3 (10.5)	Mean (SD) 67.3 (10.5)
			Median 68	Median 68
			Min, Max 42, 90	Min, Max 42, 90
Age (years)	Age (years)		Age Group - n(%)	Age Group - n(%)
40-49	40-49		40-49 80 (64.5%)	40-49 80 (64.5%)
50-59	50-59		40-54 44 (35.5%)	40-54 44 (35.5%)
60-69	60-69		50-59 32 (25.8%)	50-59 32 (25.8%)
70-79	70-79		60-69 28 (22.6%)	60-69 28 (22.6%)
80-89	80-89		70-79 25 (20.0%)	70-79 25 (20.0%)
90-99	90-99		80-89 22 (17.7%)	80-89 22 (17.7%)
100-109	100-109		90-99 20 (16.3%)	90-99 20 (16.3%)
110-119	110-119		100-109 18 (14.5%)	100-109 18 (14.5%)
120-129	120-129		110-119 16 (12.8%)	110-119 16 (12.8%)
130-139	130-139		120-129 14 (11.3%)	120-129 14 (11.3%)
140-149	140-149		130-139 12 (9.7%)	130-139 12 (9.7%)
150-159	150-159		140-149 10 (8.0%)	140-149 10 (8.0%)
160-169	160-169		150-159 8 (6.4%)	150-159 8 (6.4%)
170-179	170-179		160-169 6 (4.8%)	160-169 6 (4.8%)
180-189	180-189		170-179 4 (3.2%)	170-179 4 (3.2%)
190-199	190-199		180-189 2 (1.6%)	180-189 2 (1.6%)
200-209	200-209		190-199 1 (0.8%)	190-199 1 (0.8%)
210-219	210-219		200-209 1 (0.8%)	200-209 1 (0.8%)
220-229	220-229		210-219 1 (0.8%)	210-219 1 (0.8%)
230-239	230-239		220-229 1 (0.8%)	220-229 1 (0.8%)
240-249	240-249		230-239 1 (0.8%)	230-239 1 (0.8%)
250-259	250-259		240-249 1 (0.8%)	240-249 1 (0.8%)
260-269	260-269		250-259 1 (0.8%)	250-259 1 (0.8%)
270-279	270-279		260-269 1 (0.8%)	260-269 1 (0.8%)
280-289	280-289		270-279 1 (0.8%)	270-279 1 (0.8%)
290-299	290-299		280-289 1 (0.8%)	280-289 1 (0.8%)
300-309	300-309		290-299 1 (0.8%)	290-299 1 (0.8%)
310-319	310-319		300-309 1 (0.8%)	300-309 1 (0.8%)
320-329	320-329		310-319 1 (0.8%)	310-319 1 (0.8%)
330-339	330-339		320-329 1 (0.8%)	320-329 1 (0.8%)
340-349	340-349		330-339 1 (0.8%)	330-339 1 (0.8%)
350-359	350-359		340-349 1 (0.8%)	340-349 1 (0.8%)
360-369	360-369		350-359 1 (0.8%)	350-359 1 (0.8%)
370-379	370-379		360-369 1 (0.8%)	360-369 1 (0.8%)
380-389	380-389		370-379 1 (0.8%)	370-379 1 (0.8%)
390-399	390-399		380-389 1 (0.8%)	380-389 1 (0.8%)
400-409	400-409		390-399 1 (0.8%)	390-399 1 (0.8%)
410-419	410-419		400-409 1 (0.8%)	400-409 1 (0.8%)
420-429	420-429		410-419 1 (0.8%)	410-419 1 (0.8%)
430-439	430-439		420-429 1 (0.8%)	420-429 1 (0.8%)
440-449	440-449		430-439 1 (0.8%)	430-439 1 (0.8%)
450-459	450-459		440-449 1 (0.8%)	440-449 1 (0.8%)
460-469	460-469		450-459 1 (0.8%)	450-459 1 (0.8%)
470-479	470-479		460-469 1 (0.8%)	460-469 1 (0.8%)
480-489	480-489		470-479 1 (0.8%)	470-479 1 (0.8%)
490-499	490-499		480-489 1 (0.8%)	480-489 1 (0.8%)
500-509	500-509		490-499 1 (0.8%)	490-499 1 (0.8%)
510-519	510-519		500-509 1 (0.8%)	500-509 1 (0.8%)
520-529	520-529		510-519 1 (0.8%)	510-519 1 (0.8%)
530-539	530-539		520-529 1 (0.8%)	520-529 1 (0.8%)
540-549	540-549		530-539 1 (0.8%)	530-539 1 (0.8%)
550-559	550-559		540-549 1 (0.8%)	540-549 1 (0.8%)
560-569	560-569		550-	



Categories of Adverse Events (Safety Analysis Set)

	Number of patients ^a	D=Guinea Pig (N=100)	Placebo=Guinea Pig (N=100)
DOC			
7/200	12 (3.0)	14 (3.0)	
Possibly related to any study medication ^b	2 (0.6)	3 (0.3)	
Any SAE (including SAE with serious or fatal)	100 (27.5)	100 (0.0)	
Possibly related to any study medication ^b	13 (3.0)	13 (3.0)	
Any AE leading to discontinuation of study treatment			
Any	44 (31.0)	55 (32.0)	
Discontinued on placebo	28 (3.3)	38 (3.3)	
One and/or C ^b	40 (3.3)	47 (3.3)	
Possibly related to any study medication ^b	10 (3.0)	10 (3.0)	
Any AE leading to dose interruption/hold or reduction			
Any	22 (20.0)	24 (23.0)	
Discontinued on placebo	22 (20.0)	13 (24.0)	
One and/or C ^b	22 (20.0)	24 (23.0)	
Any SAE ^b	40 (32.0)	39 (32.0)	
Possibly related to any study medication ^b	38 (32.0)	34 (32.0)	
Any adverse reaction SAE ^b	13 (3.0)	9 (3.0)	
Any other significant SAE ^b	28 (30.0)	29 (30.0)	
Unknown SAEs ^b	14 (27.0)	14 (27.0)	
Other SAEs ^b	79 (26.0)	77 (27.0)	
Nonserious events (common SAEs SAEs)	276 (79.0)	280 (77.0)	

	A	B	C	D
Any AE with serious or fatal			10 (3.0)	13 (3.0)
Possibly related to any study medication ^b			2 (0.6)	3 (0.3)
Any SAE (including SAE with serious or fatal)			100 (27.5)	100 (0.0)
Possibly related to any study medication ^b			13 (3.0)	13 (3.0)
Any AE leading to discontinuation of study treatment				
Any			44 (31.0)	55 (32.0)
Discontinued on placebo			28 (3.3)	38 (3.3)
Dose and/or C ^b			40 (3.3)	47 (3.3)
Possibly related to any study medication ^b			10 (3.0)	10 (3.0)
Any AE leading to dose interruption/hold or reduction				
Any			22 (20.0)	24 (23.0)
Discontinued on placebo			22 (20.0)	13 (24.0)
Dose and/or C ^b			22 (20.0)	24 (23.0)
Possibly related to any study medication ^b			38 (32.0)	34 (32.0)
Any adverse reaction SAE ^b			13 (3.0)	9 (3.0)
Any other significant SAE ^b			28 (30.0)	29 (30.0)
Unknown SAEs ^b			14 (27.0)	14 (27.0)
Other SAEs ^b			79 (26.0)	77 (27.0)
Nonserious events (common SAEs SAEs)			276 (79.0)	280 (77.0)

Option B: Ad-hoc programming - rtf/xml/txt parser

[illegible]

```
<?xml version='1.0' encoding='UTF-8'?><br>  
    <!--XML language --><br>  
    <xhtml:body xmlns:xhtml="http://www.w3.org/1999/xhtml"><br>  
        </xhtml:body></xhtml:html>
```

Option C: Reproduce the output via SAS programming

SAS programming



2) Text Spec to Code Translation

LBORRES = For RAW.LB records:
 If RAW.LB.MODULE='LB3' and RAW.LB.LBVDIP not missing then set to upcase(vvalue(RAW.LB.LBVDIP)), else if
 RAW.LB.MODULE='LB3' and RAW.LB.LBORRES not missing then set to RAW.LB.LBORRES, else if RAW.LB.MODULE='LB4'
 and RAW.LB.LBVDIP not missing then set to upcase(vvalue(RAW.LB.LBVDIP)), else if RAW.LB.MODULE='LB4' and
 RAW.LB.LBORRES not missing then set to RAW.LB.LBORRES, else if RAW.LB.MODULE='LB6' and RAW.LB.LBVDIP not
 missing then set to upcase(vvalue(RAW.LB.LBVDIP)), else if RAW.LB.MODULE='LB6' and RAW.LB.LBORRES not missing then
 set to RAW.LB.LBORRES, else set to RAW.LB.LBORRES.

[illegible]

```

STUDYID=TEST001 ;
DOMAIN='OH' ;
USUBJID=compress (put (substr (STUDYID, 4, 3), 3, 3) || '-' || put (SITEID, 3, 3) || put (patno, 3, 3)) ;
SUBJID=concat (put (siteid, 3, 3), '-', put (patno, 3, 3)) ;
SITEID=compress (put (siteid, best, 3)) ;
INVID=invno ;
INVMAN=PUT (invname, INVV) ;

if nmisc (exbegdat, exbegtim)=0 then RFSTDTCT=put (dhms (exbegdat , 0,0, exbegtim ), is601dtt);
if nmisc (exenddat, exendtim)=0 then RFENDTCT=put (dhms (exenddat , 0,0, exendtim ), is601dtt);
if nmisc (birthdt, addmat)=0 then RFBRTHCT=PUT (birthdt, is601dtt);
if nmisc (birthdt, addmat)=0 then RFBRTHCT=PUT (birthdt, is601dtt);
AGE=floor ((intck ('month', birthdt, addmat ) - (day (addmat ) < day (birthdt))) / 12) ;
AGEU='YEARS' ;
SEX=substr (put (sexm, sex, ), 1, 1) ;
RACE=put (race, race, ) ;
ETHNIC=' ' ;

if trt='A' then ARMCD='TEST';
else if trt='B' then ARMCD='PLACEBO';
else ARMCD='';
addmat=put (trt, test, );
COUNTRY='USA';
if nmisc (exbegdat, addmat)=0 then do;
if ex-
end;
else
proc sort data=ocdata;demog out=demog ;
by studyid ptno ;
run;

/*Combining all the data into one dataset*/

data combine_all ;
merge random (in=a) invinfo (in=b) demog (in=c) ;
by studyid ptno ;
run;

/*Creating Final OM dataset with required attributes*/

data edtm.dml;
set combine_all ;

attrib STUDYID length = 540 label = 'Study Identifier';
attrib DOMAIN length = 58 label = 'Domain Abbreviation';
attrib USUBJID length = 540 label = 'Unique Subject Identifier';
attrib SUBJID length = 540 label = 'Subject Identifier for the Study';
attrib RFSTDTCT length = 564 label = 'Study Reference Start Date/Time';
attrib RFENDTCT length = 564 label = 'Subject Reference End Date/Time';
attrib SITEID length = 540 label = 'Study Site Identifier';
attrib DATA length = 540 label = 'Investigator Identifier';
attrib INVMAN length = 540 label = 'Investigator Name';
attrib BRTHCT length = 564 label = 'Date/Time of Birth';
attrib AGE length = 8 label = 'Age in AGU at RFSTDTCT';
attrib AGEU length = 210 label = 'Age Units';
attrib SEX length = 52 label = 'Sex';
attrib RACE length = 540 label = 'Race';
attrib ETHNIC length = 540 label = 'Ethnicity';
attrib ARMCD length = 520 label = 'Planned Arm Code';
attrib ARM length = 540 label = 'Description of Planned Arm';
attrib COUNTRY length = 53 label = 'Country';
attrib DMY length = 8 label = 'Study Day of Collection';

keep STUDID DOMAIN USUBJID SUBJID RFSTDTCT RFENDTCT SITEID INVID INVMAN
BRTHCT AGE AGEU SEX RACE ETHNIC ARMCD ARM COUNTRY DMY ;

```



Problem examples

Manual coding?



2) Text Spec to Code Translation

```
1 ****
2 * Program name ..... : AE.sas
3 * Type ..... : SDTM*
4 * Purpose ..... : Creation of sdsm.AE and sdsm.suppAE
5 * Author ..... : XXXX
6 * Datetime created ..... : 07/13/2022, 12:53:15
7 * Input datasets ..... : [NOTE: Complete as needed from the mapping specification]
8 * Extract Rule ..... : [where RAW.AE.AEANY="1"]
9 * Macros used ..... : %setup, %localsetup
10 * Output files ..... : sdsm.AE, sdsm.suppAE
11 * Revision history ..... :
12
13 * Date ..... Author ..... Ref ..... Revision
14 ****
15
16 ****
17 ** Setup calls ..... :
18 ****
19 %autoexec;
20
21 %let sdsm_domain = AE;
22
23 %let AESort_Order = STUDYID USUBJID AESTDTC AEDECOD;
24
25 %let kname = STUDYID DOMAIN USUBJID AESQO AESPID AETERM AELLT AELLTC AEDECOD AEPTCD AEHLT AEHLTQ
26 AESDTH AESHOSP AESLIFE AESMIE AECONTRT EPOCH AESTDTC AEENDTC AESTDY AEENDY;
27
28 %preprocess(domain=AE);
29
30
31 data RAW_AE;
32
33   .attrib
34     >>> STUDYID ..... label='Study Identifier' ..... length=421
35     ...
36     >>> AEREL ..... label='Causality' ..... length=200
37     >>> AEOUT ..... label='Outcome of Adverse Event' ..... length=200
38
39   .set
40     >>> RAW.AE(rename=(AESER = _AESER.) where=(AEANY="1"));
41
42     >>> /* Set to "AE-" || left(trim(put(RAW.AE.AENO,best.))) if RAW.AE.AENO is not missing. >>> */
43     >>> if ^missing(AENO) then do; AESPID = "AE-" || left(trim(put(AENO,best.))); end;
44
45     >>> /* Set to RAW.AE.SOC_CODE */
46     >>> AESOCCD = SOC_CODE;
47
48     >>> /* Assign RAW.AE.AESEVMAX as format. */
49     >>> AESEV=
50       >>> ifc(AESEVMAX = "1", "MILD",
51       >>> ifc(AESEVMAX = "2", "MODERATE",
52       >>> ifc(AESEVMAX = "3", "SEVERE",
53       >>> AESEVMAX)) || ;
54     >>> /* Assign RAW.AE.AEDECOD as format. */
```

Program header

System macro variable

Attributes, good format

Rename, Where

Codelist

```
330 /* ===== temp ===== */
331 data temp;
332   merge RAW_AE(in=a) RAW_SERAE;
333   by SUBJECT AENO;
334   if a
335     >>> DSSPID=strip(put(DSASPID1,best.));
336     >>> DSTERM="MAIN INFORMED CONSENT 2";
337     >>> DSDECOD="INFORMED CONSENT 2 OBTAINED";
338     >>> DSCAT="PROTOCOL MILESTONE";
339     >>> DSDTC=VIS_DAT;
340     >>> DSSTDTC=DSSTDTC2;
341     >>> VISITDTC=VIS_DAT;
342   output;
343   end;
344
345 data temp;
346
347   .attrib
348     >>> STUDYID ..... label='Study Identifier' ..... length=421
349     >>> DOMAIN ..... label='Domain Abbreviation' ..... length=421
350     >>> .....
351
352   /* Compare VSDTC with SE.SESTDTC and SE.SEENDTC per subject per EPOCH. Set as 'SCREENING' if SE.SESTDTC < QSDTC < SE
353   < SE.SEENDTC where SE.EPOCH= TREATMENT. Set as 'FOLLOW-UP' if SE.SESTDTC < QSDTC < SE.SEENDTC where SE.EPOCH= FOLLOW
354   if SE.SESTDTC and SEENDTC include time part, then we need to compare the time, if no, only to compare date part. */
355   /* Variable EPOCH tod - CANNOT PARSE
356   EPOCH=Compare VSDTC with SE.SESTDTC and SE.SEENDTC per subject per EPOCH.
357   Set as 'SCREENING' if SE.SESTDTC < QSDTC < SE.SEENDTC where SE.EPOCH=SCREENING; Set as 'TREATMENT' if SE.SESTDTC <
358   /* Variable EPOCH tod */;
359
360   >>> STUDYID = "00000000";
361   >>> STUDYID = "00000000";
362   >>> "AE"
363   >>> DOMAIN = "AE";
364   >>> /* Concatenate STUDY and SUBJECT separating by '/'
365   USUBJID = cat(strip(STUDY), '/', strip(SUBJECT));
366
367   /* Calculation of AESTDY = (Numeric version of date part of AESTDTC - Numeric version of date part of DM.RFSSTDTC) ... Note1: Add 1 if cal
368   then AESTDY is equal to missing.
369   %letm_dy(INVAR_DTC = AEstdct, OUTVAR_DY = AEstdy);
370
371   /* Calculation of AEENDY = (Numeric version of date part of AEENDTC - Numeric version of date part of DM.RFSSTDTC) ... Note1: Add 1 if cal
372   then AEENDY is equal to missing.
373   %letm_dy(INVAR_DTC = AEndct, OUTVAR_DY = AEndy);
374
375   run;
376
377   /* Compare AESTDTC with SE.SESTDTC and SE.SEENDTC per subject per EPOCH. Set as 'SCREENING' if SE.SESTDTC < AESTDTC < SE.SEENDTC where
378   < SE.SEENDTC where SE.EPOCH= TREATMENT; Set as 'FOLLOW-UP' if SE.SESTDTC < AESTDTC < SE.SEENDTC where SE.EPOCH= FOLLOW-UP; If date part
379   Note: If SE.SESTDTC and SEENDTC include time part, then we need to compare the time, if no, only to compare date part.
380   %letm_epoch(indata=final1, outdata=final2, domain=AE, debug=Y);
```

Data Join/Combine

Grouping process

Todo placeholder

Variable level

Data level



New Perspective

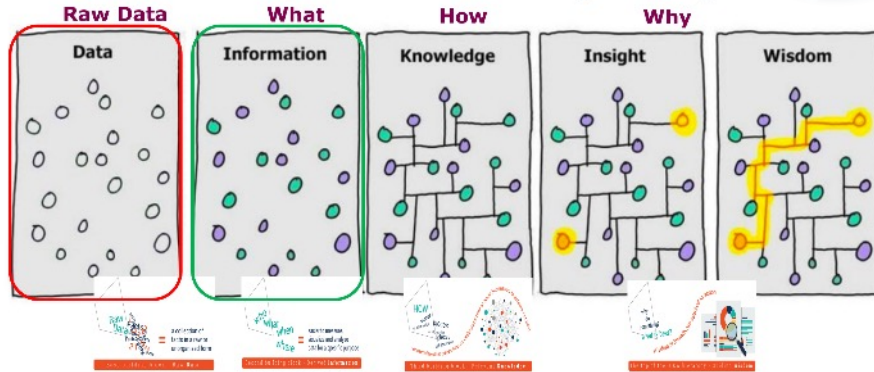


From Unstructured Data to Structured Data – AI Solution

Human-driven to Software-driven



DIKW Model (Data-to-Information-to-Knowledge-to-Wisdom)



Unstructured Data

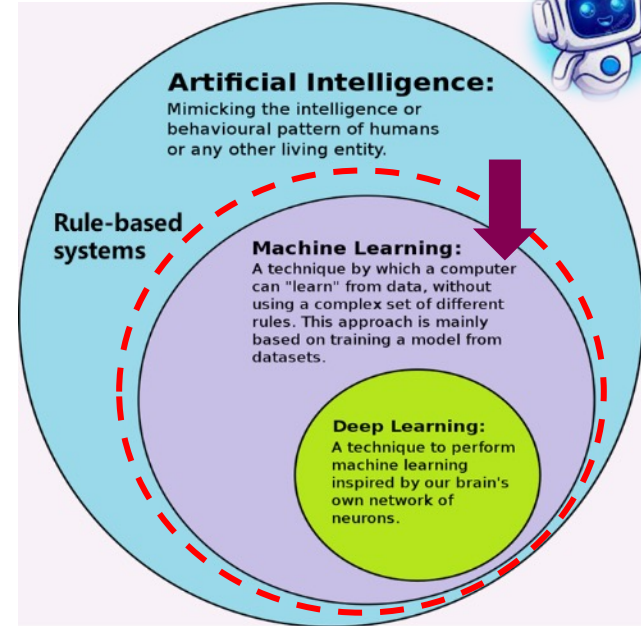
The university has 5600 students. Shaun (ID Number: 160801), 18 years old Communication study. Linh with ID number 160802, majoring in Accounting and is 20 years old. Ahmed from Psychology study program, 19 years old, ID number 160803.

Semi-Structured Data

```
<University>
<ID Number="160801">
  <Name="Shaun">
    <Age="18">
      <Program="Communication">
        <ID Number="160802">
          <Name="Linh">
            <Age="20">
              <Program="Accounting">
                ..... </University>
```

Structured Data

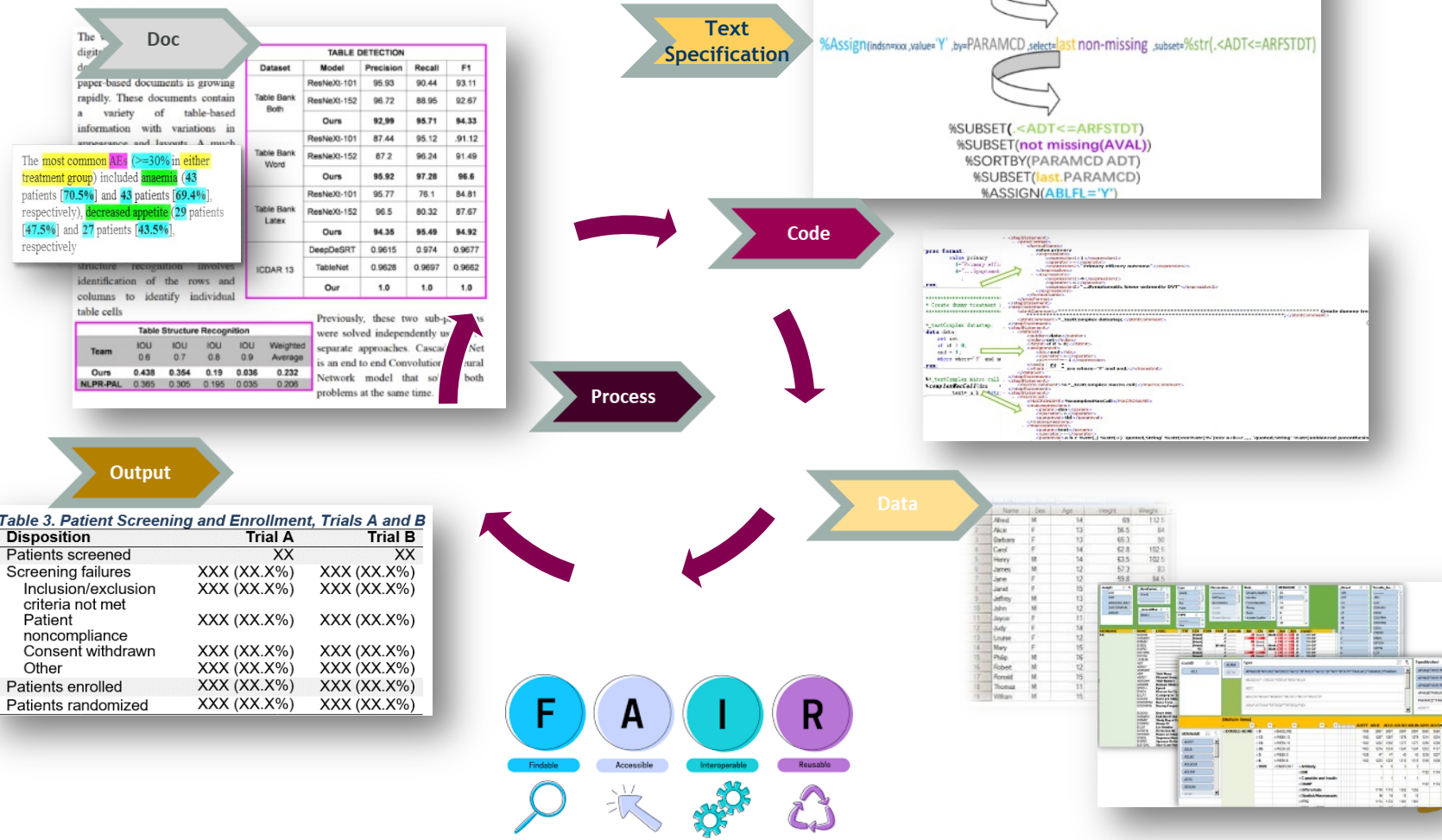
ID	Name	Age	Program
160801	Shaun	18	Communication
160802	Linh	20	Accounting
160803	Ahmed	19	Psychology



How deep learning is a subset of machine learning and how machine learning is a subset of artificial intelligence (AI) https://en.wikipedia.org/wiki/Deep_learning



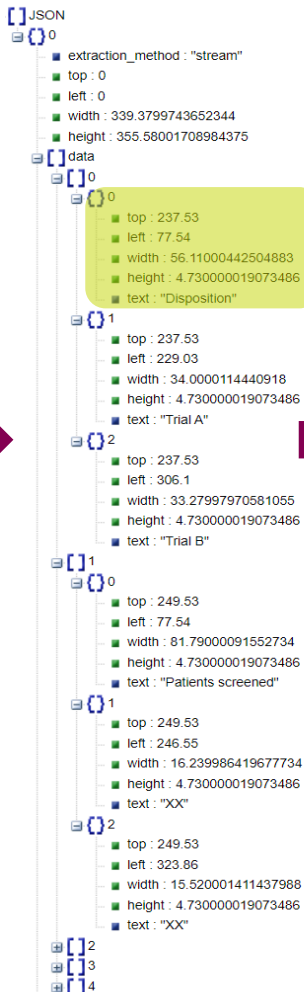
Digitalization for the Trial Artifacts



Solutions

1) Output Table analysis

Disposition	Trial A	Trial B
Patients screened	XX	XX
Screening failures	XXX (XX.X%)	XXX (XX.X%)
Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
Other	XXX (XX.X%)	XXX (XX.X%)
Patients enrolled	XXX (XX.X%)	XXX (XX.X%)
Patients randomized	XXX (XX.X%)	XXX (XX.X%)



	A	B	C	D	E
1	rowNum	Lev0	Lev1	V2	V3
2	1	Disposition		Trial A	Trial B
3	2	Patients screened		XX	XX
4	3	Screening failures		XXX (XX.X%)	XXX (XX.X%)
5	4	Screening failures	Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
6	5	Screening failures	Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
7	6	Screening failures	Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
8	7	Screening failures	Other	XXX (XX.X%)	XXX (XX.X%)
9	8	Patients enrolled		XXX (XX.X%)	XXX (XX.X%)
10	9	Patients randomized		XXX (XX.X%)	XXX (XX.X%)

Challenges

- Rows/Columns
- Nested Levels
- Paragraphs/Pages
- Dynamic table area
- Broken lines
- Dynamic spacing
- ...



Solutions

2) Text Spec to Code Translation

Text Spec (human language + code piece)

LBORRES = For RAW.LB records:

If RAW.LB.MODULE='LB3' and RAW.LB.LBVDIP not missing then set to upcase(vvalue(RAW.LB.LBVDIP)), else if RAW.LB.MODULE='LB3' and RAW.LB.LBORRES not missing then set to RAW.LB.LBORRES, else if RAW.LB.MODULE='LB4' and RAW.LB.LBVDIP not missing then set to upcase(vvalue(RAW.LB.LBVDIP)), else if RAW.LB.MODULE='LB4' and RAW.LB.LBORRES not missing then set to RAW.LB.LBORRES, else if RAW.LB.MODULE='LB6' and RAW.LB.LBVDIP not missing then set to upcase(vvalue(RAW.LB.LBVDIP)), else if RAW.LB.MODULE='LB6' and RAW.LB.LBORRES not missing then set to RAW.LB.LBORRES, else set to RAW.LB.LBORRES.

If RAW.LB.MODULE='LB3' FORMULA and RAW.LB.LBVDIP VAR not missing then set to upcase(vvalue(RAW.LB.LBVDIP)) FUNC, else if RAW.LB.MODULE='LB3' FORMULA and RAW.LB.LBORRES VAR not missing then set to RAW.LB.LBORRES VAR, else if RAW.LB.MODULE='LB4' FORMULA and RAW.LB.LBVDIP VAR not missing then set to upcase(vvalue(RAW.LB.LBVDIP)) FUNC, else if RAW.LB.MODULE='LB4' FORMULA and RAW.LB.LBORRES VAR not missing then set to RAW.LB.LBORRES VAR, else if RAW.LB.MODULE='LB6' FORMULA and RAW.LB.LBVDIP VAR not missing then set to upcase(vvalue(RAW.LB.LBVDIP)) FUNC, else if RAW.LB.MODULE='LB6' FORMULA and RAW.LB.LBORRES VAR not missing then set to RAW.LB.LBORRES VAR, else set to RAW.LB.LBORRES VAR.

```
LBORRES= ifc(MODULE='LB3' and ^missing(LBVDIP), upcase(vvalue(LBVDIP)),
            ifc(MODULE='LB3' and ^missing(LBORRES), LBORRES,
            ifc(MODULE='LB4' and ^missing(LBVDIP), upcase(vvalue(LBVDIP)),
            ifc(MODULE='LB4' and ^missing(LBORRES), LBORRES,
            ifc(MODULE='LB6' and ^missing(LBVDIP), upcase(vvalue(LBVDIP)),
            ifc(MODULE='LB6' and ^missing(LBORRES), LBORRES,
            LBORRES )))))))::;
```

SAS programs
(or R/Python/SQL)



- **Case Sharing**

- 1) **Output Table Analysis**

- Side by Side Table Generation

- 2) **Text Spec to Code Translation**

- SDTM Mapping Code Generation



Side by Side Table Generation

- Business needs

Needs for clinical document writing

- Side-by-side tables (Global vs Asia vs China populations) are needed for China submission

SAS programming not the best solution

- Additional resource/budget/time needed
- Start from Scratch/Format Adjustment

Manual creation not the best solution

- Time consuming
- Pron to error
- Intensive QC required



Side by Side Table Generation

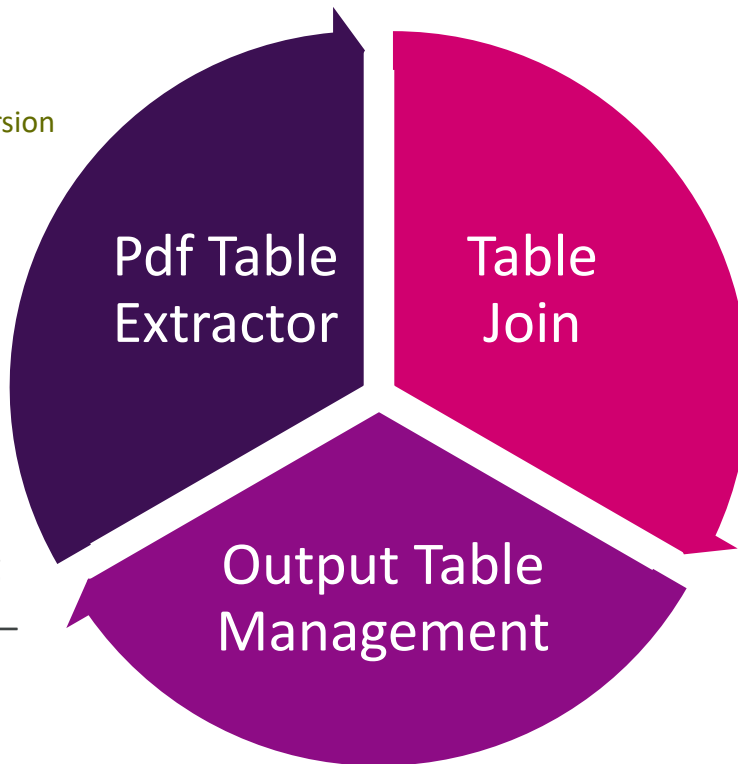
- Current key functions

Extract data from TFL files:

- Legacy TFLs without editable version
- Inconsistent format of tables
- Table across multiple pages

Benefits

- 3-5 minutes/table
 - Spotted QC is still required
- Batch and Real-time processing
- Data extraction accuracy 100% – No OCR(Optical character recognition)



Join table:

- Join by auto-populated/flexible keys
- 2+ tables join

Manage the outlook of a table:

- Filter content or data
- Sort values
- Other calculations users may need



Side by Side Table Generation

- Process

PDF table



DOC(X)/RTF table



Table
Extract



Table
Match/Join



Table
Filter/Format



PDF Table Extractor

Table Join

Output Table

Use this tool to extract structured, machine-readable tables from PDF reports in a few clicks. Load a PDF into the app and extract tables right in the browser, ready for export to Excel

Upload PDF file

	V1	V2	V3	V4	V5
1					



SDTM Mapping Code Generation

- Business needs

SDTM spec and programming codes consistency

- Updates should be taken in both files.

SDTM programming code not reusable

- Reference programming code should be checked line by line before implementation.

Manual creation not the best solution

- Time consuming
- Pron to error
- Intensive QC required for simple and standardized variable generation



SDTM Mapping Code Generation

- **Designed feature**
 - **Natural Language Spec -> Metadata -> code**
 - **Spec adaptive to different standard (across TA/vendor)**
 - **70/30% -> 80/20%**
 - **Standard algorithm accumulated**
 - **Rerun applicable with updated user-defined code**



SDTM Mapping Code Generation

• Text Spec to Code

• Minimum input from Spec

• Type I (from Spec) – simple derivation

- Copy/Drop/Rename
- Assign - function call

• Type II (from Spec) – standard derivation

- Standard Mapping macro call (top down)
- User-defined Pattern macro call, some chance to reuse (bottom up)

• Type III (from open code) – complex derivation

- User-defined complex algorithm, little chance to reuse
- Do it in open code
- Rerun applicable with updated Code & Spec

```

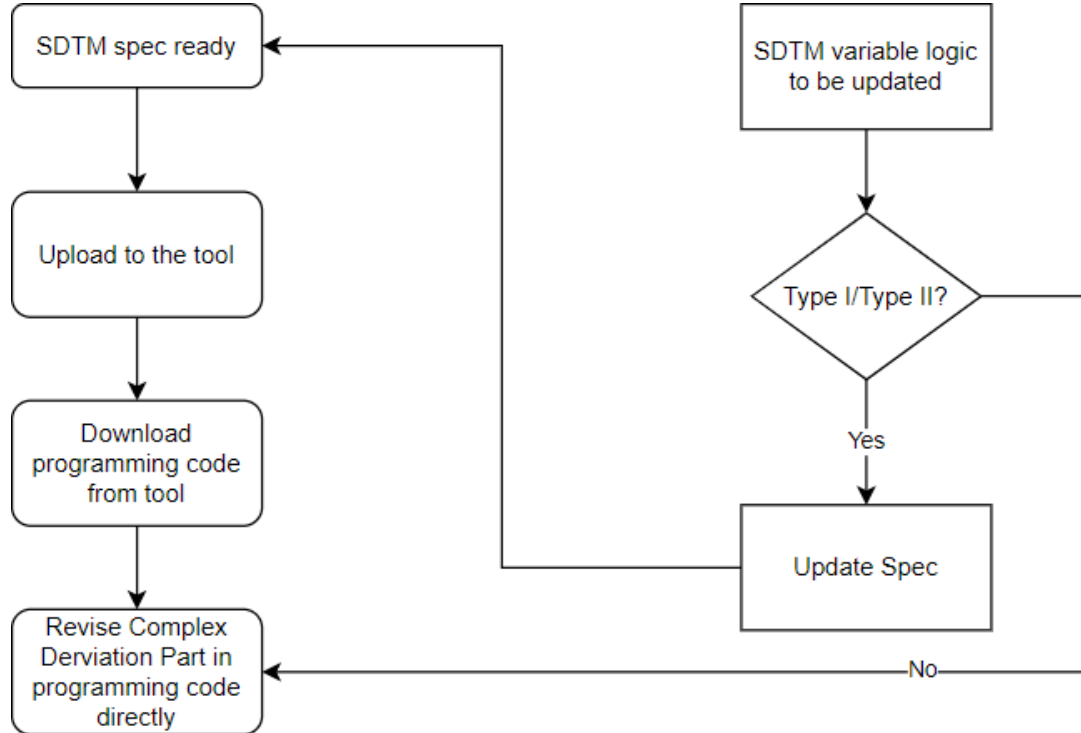
206 _domain='AE';
207 _module='AE SERAE';
208 _aedw=put( aedis, $NY.);
209 _aeendtc=%dtc(INVAR_DAT= ae.aeendat ,INVAR_TIM= ae.aeentim ,OUTVAR_DTC= ae._aeendtc );
210 _aeout=upcase(putc( aeout,vformat(aeout)));
211 _aerel=put( aerel, $AERELF.);
212 _aeser=upcase(putc( aeser,vformat(aeser)));
213 _aesev=upcase(putc( aesevmax,vformat(aesevmax)));
214 _aestdtc=%dtc(INVAR_DAT= ae.aestdat ,INVAR_TIM= ae.aesttim ,OUTVAR_DTC= ae._aestdtc );
215 _aeterm=aeterm;
216 _aehlgtcd=hltcode;
217 _aehlgt=hltname;
218 _aehltd=hlt_code;
219 _aehlthlt_name;
220 _ip=ip;
221 run;
222
223 *****
224 * Execute standalone macros
225 *****
226 %dy(INVAR_REFDTC=&_templib._dm.rfstdtc ,INVAR_DTC=&_templib._ae.aeendtc ,OUTVAR_DY=&_templib._ae.aeendy ); ;
227 %dy(INVAR_REFDTC=&_templib._dm.rfstdtc ,INVAR_DTC=&_templib._ae.aestdtc ,OUTVAR_DY=&_templib._ae.aestdy ); ;
228 %seq(DATASET_LAYOUT=STUDYID USUBJID AESTDTC AEDECOD ,OUTVAR_SEQ=&_templib._ae.aeseq); ;
229
4  ***For RAW.PREG records:\n'if RAW.PREG.MODULE='PRE...***;
5  /* variable LBSPEC code */
6  VISITNUM=visit;
7  $sdm dlc(INVAR_DAT=samp dat,INVAR_TIM=,OUTVAR_DTC=$sdm domain.dlc );
8  if vvalue(ASM_APU)="N" then LBSTAT="NOT DONE";
9  if upcase(module)="PREG" then
10 do;
11 LBORRES=upcase(vvalue(PREGSTU));
12 LBSPEC="URINE";
13 end;
14 if upcase(module)="PREG1" then
15 do;
16 LBORRES=upcase(vvalue(PREGSTS));
17 LBSPEC="SERUM";
18 end;
19 LBTESTCD="HCG";
20 LBTEST="Chorioqonadotropin Beta";
21 LBSTAT="PREGNANCY TEST";
22 /* LBSPEC= */
23 /* nan nan nan nan
24 /* variable LBSPEC code */
25 run;

```



SDTM Mapping Code Generation

- **Process***



SDTM Mapping Code Generation

- Training Summary in first stage

Training Set	Total Variable Number	Automated Variable Number (Type I)	Automated Variable Number (Type II)	Actual Automation Rate (Type I + Type II)	Potential Automation Rate
AE	62	37(59.6%)	16(25.8%)	53(85.4%)	62(100%)
VS	26	14(53.8%)	8(30.8%)	22(84.6%)	24(100%)
EC	38	22(57.9%)	5(13.2%)	27(71.1%)	36(94.7%)
DM	31	14(45.1%)	4(12.9%)	18(58.0%)	26(83.8%)



https://rstudio.scp.astrazeneca.net/s/aef2cc391f1239bee297a/p/48ecb705/ | Open in Browser

Mapping Spec File

Code Generated

Generation Process

Upload Mapping Spec File

Browse...

No file selected

Name of Sheet (Case-Sensitive)

☐ use template SAS code?



Summary



Summary

- New technology is an evitable trend in pharma industry
 - General solution for unstructured data digitization
 - Save human effort to avoid low-value work
 - Combination of Machine and Human(AI+HI)
 - From Rule-based solution to Machine Learning solution
- Business always pursuits efficiency and shorten timeline
 - Side by Side Table Generation Tool accelerates post DBL submission
 - SDTM Mapping Code Generation empowers the SDTM creation





Acknowledgements

Sponsor: Meng Chen

Programming: Xinya Wang, Zhanjian Dong, Kang Zhao, Fan Zhang, Zhilong Qi

Medical Communication: Ning Zheng, Miao Tong, Simei Yu, Qingdan Du

Data Scientist Intern: Wenkai Xiang, Jessica Chen, Hengjie Cui

*Thank
you*

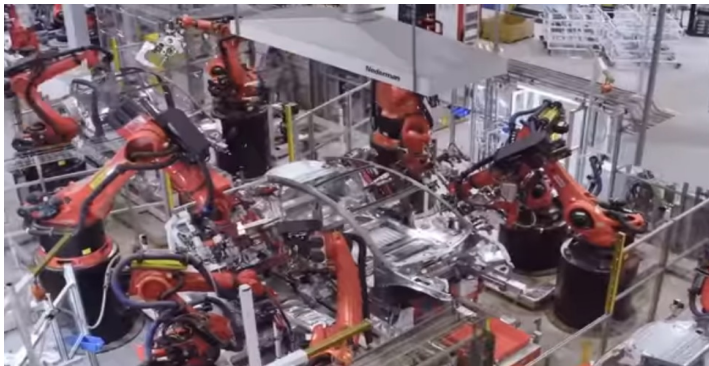


backup



Our Problem

Automobile Manufacturing Process



Tesla Gigafactory 3 Shanghai

<https://www.youtube.com/watch?v=-Ds1xV7M2gI>

- 100% repeatable process
- Machine is the center
- Defective product destroyed



or

Transportation System



Pony.ai self-drive

<https://www.youtube.com/watch?v=iKFV7qjWaD4>

- Always change
- Human in the loop
- Corner case matters



• .rtf/.doc/.docx -> .pdf

PDF Table Extractor Table Join Output Table

Use this tool to extract structured, machine-readable tables from PDF reports in a few clicks. Load a PDF into the app and extract tables right in the browser, ready for export to Excel

Upload PDF file FDA-2022-N-1961-0002

Add suffix to download file

Download

1

☒ Table Detect

☒ Level Hierarchy

Scrape

Go!

Cancel

	tableID	tableNum	pageNum	rowNum	Lev0	Lev1	V2	V3
1	no tableID detected	1	1	1	Patients screened		XX	XX
2	no tableID detected	1	1	2	Screening failures		XXX (XX.X%)	XXX (XX.X%)
3	no tableID detected	1	1	3	Screening failures	Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
4	no tableID detected	1	1	5	Screening failures	Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
5	no tableID detected	1	1	7	Screening failures	Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
6	no tableID detected	1	1	8	Screening failures	Other	XXX (XX.X%)	XXX (XX.X%)
7	no tableID detected	1	1	9	Patients enrolled		XXX (XX.X%)	XXX (XX.X%)
8	no tableID detected	1	1	10	Patients randomized		XXX (XX.X%)	XXX (XX.X%)

Table 3 provides an example of how screening and enrollment data should be displayed. The populations should be defined as follows:

- Screened population: All patients screened for entry into the trial
- Enrolled population: All patients who signed a consent form for participation in the trial
- Randomized population: All patients randomized to a treatment arm; often referred to as the intention-to-treat population
- If screening failure data are available, provide reasons for screen failure, such as "patient noncompliance," "consent withdrawn," "inclusion/exclusion criteria not met," "other." Request breakdown of "inclusion/exclusion criteria not met" and "other," because understanding reasons for screen failure can be informative to assess the generalizability of the trial.

Table 3. Patient Screening and Enrollment, Trials A and B

Disposition	Trial A	Trial B
Patients screened	XX	XX
Screening failures	XXX (XX.X%)	XXX (XX.X%)
Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
Other	XXX (XX.X%)	XXX (XX.X%)
Patients enrolled	XXX (XX.X%)	XXX (XX.X%)
Patients randomized	XXX (XX.X%)	XXX (XX.X%)

Source: [include Applicant source, datasets and/or software tools used].

Table 4 is an example table showing all reasons reported in the study disposition dataset that should be included. Consider the following study populations:

- Modified intention-to-treat (mITT) population: Subset of the ITT population allowing the exclusion of some randomized patients in a justified way (e.g., patients who were deemed ineligible after randomization or certain patients who never started treatment)
- Safety population: All patients considered in the safety analyses who received at least one dose of the study drug in the trials submitted
- Per-protocol population: Only those patients who completed the treatment originally allocated and planned without major protocol violations

<https://www.regulations.gov/document/FDA-2022-N-1961-0002>



- .lst -> .pdf

PDF Table Extractor

Table Join

Output Table

Use this tool to extract structured, machine-readable tables from PDF reports in a few clicks. Load a PDF into the app and extract tables right in the browser, ready for export to Excel

Upload PDF file

table layout test lst.pdf

Add suffix to download file

Download

1

☐ Table Detect☒ Level Hierarchy

Scrape

	tableID	tableNum	pageNum	rowNum	Lev0	Lev1	V2	V3
1	no tableID detected	1	1	1	Patients screened		XX	XX
2	no tableID detected	1	1	2	Screening failures		XXX (XX.X%)	XXX (XX.X%)
3	no tableID detected	1	1	3	Screening failures	Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
4	no tableID detected	1	1	5	Screening failures	Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
5	no tableID detected	1	1	6	Screening failures	Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
6	no tableID detected	1	1	7	Screening failures	Other	XXX (XX.X%)	XXX (XX.X%)
7	no tableID detected	1	1	8	Patients enrolled		XXX (XX.X%)	XXX (XX.X%)
8	no tableID detected	1	1	9	Patients randomized		XXX (XX.X%)	XXX (XX.X%)

table layout test.lst - Notepad

1 / 1

80%

Table 3 provides an example of how screening and enrollment data should be displayed. The populations should be defined as follows:

- Screened population: All patients screened for entry into the trial
- Enrolled population: All patients who signed a consent form for participation in the trial
- Randomized population: All patients randomized to a treatment arm; often referred to as the intention-to-treat population
- If screening failure data are available, provide reasons for screen failure, such as "patient noncompliance," "consent withdrawn," "inclusion/exclusion criteria not met," "other." Request breakdown of "inclusion/exclusion criteria not met" and "other," because understanding reasons for screen failure can be informative to assess the generalizability of the trial.

Table 3. Patient Screening and Enrollment, Trials A and B

Disposition	Trial A	Trial B
Patients screened	XX	XX
Screening failures	XXX (XX.X%)	XXX (XX.X%)
Inclusion/exclusion criteria not met	XXX (XX.X%)	XXX (XX.X%)
Patient noncompliance	XXX (XX.X%)	XXX (XX.X%)
Consent withdrawn	XXX (XX.X%)	XXX (XX.X%)
Other	XXX (XX.X%)	XXX (XX.X%)
Patients enrolled	XXX (XX.X%)	XXX (XX.X%)
Patients randomized	XXX (XX.X%)	XXX (XX.X%)

Source: [include Applicant source, datasets and/or software tools used].

Table 4 is an example table showing all reasons reported in the study disposition dataset that should be included. Consider the following study populations:

