

Split Character Variables into Meaningful Text

BY
Savithri Jajam
Vamshi Matta
Lavanya Peddibhotla

June 14th – 18th , 2021

Copyright © 2021 Covance Inc. All Rights Reserved.



COVANCE
by labcorp

Savithri Jajam Biography

- Savithri Jajam holds a Master's degree from Western Illinois University where she studied Masters in chemistry.
- She is currently working at Covance as Senior Principal statistical programmer.
- She has extensive CDISC experience, working with both SDTM and ADaM standards.

Covance by Labcorp Company Descriptions:

50-Word Company Description

Covance by Labcorp is a leading global life sciences company, which provides contract research services to the drug, medical device and diagnostics, crop protection and chemical industries. Employing over 26,000 people worldwide, we provide comprehensive drug development solutions to advance health and power clear, confident decisions.

www.covance.com.

100-Word Company Description

Covance by Labcorp is a leading global life sciences company, which provides contract research services to the drug, medical device and diagnostics, crop protection and chemical industries. Employing over 26,000 people worldwide, we provide comprehensive drug development solutions and are on a mission to advance health and power clear, confident decisions.

Covance is a global leader in nonclinical safety assessment, clinical trial testing and clinical trial management services. Our unique perspectives are based on decades of scientific, medical, and regulatory expertise. Together with our clients, Covance by Labcorp supports the development of innovative, life-changing treatments. Visit us at www.covance.com.

Introduction:

- Most of the submission datasets must follow CDISC guidelines in order to go for a submission.
- One of the primary criteria is to have 200 lengths for any character variable.
- To follow this is, when we have data of more than 200 and we must split it into multiple variables with meaningful text.

Introduction Cont:

Some of the challenges:

- **Example 1:** we don't know how lengthy the text is and how many variables it must split.
- **Example 2:** we are sure is multiple datasets had lengthy texts. Like Device, Deviation or Comments domains.
- **Example 3:** we need multiple methods to be used in prod and validation as we need Independent validation.

Simple Method: 1

```
data qc1;
set source.dv;
  describe_deviation=strip(compbl(describe_deviation));
  length=length(describe_deviation);
  c=substr(describe_deviation,201,1);
  if .<length<=200 and missing(c) then dvterm=describe_deviation;
  else if 200<length<=400 or not missing(c) then do;
    if substr(describe_deviation,200,1)="" or substr(describe_deviation,201,1)="" then do;
      dvterm=substr(describe_deviation,1,200);
      dvterm2=strip(substr(describe_deviation,201));      end;
    else do;
      length1=200-length(scan(substr(describe_deviation,1,200),-1,""));
      dvterm=substr(describe_deviation,1,length1);
      dvterm2=strip(substr(describe_deviation,length1+1,200));      end;
    end;
  end;
run;
```

Simple Method: 1

⚠ DVTERM	⚠ dvterm2	⚠ Describe_Deviation	Ⓜ length
[Day 43] The subject'...	(13May2017 through...	[Day 43] The subject's...	224
Blood transfusion his...	document with and ...	Blood transfusion hist...	248
[0044-301A] Day 18...	was informed in adv...	[0044-301A] Day 183 ...	224
[0045-301A]Day71- ...		[0045-301A]Day71- T...	172
[0045-301A] Day71- ...		[0045-301A] Day71- ...	129
[0045-302A] Day 57-...		[0045-302A] Day 57- ...	157

Simple Method: 2

```
data adcm;  
    set cm;  
    len=length(Level_5_Text);  
    length orig $400. cmdecod cmdecod1 $200;  
        orig=Level_5_Text;  
    pos=findc(orig, ' ', -201);  
    cmdecod=strip(substr(orig, 1, pos));  
    cmdecod1=strip(substr(orig, pos+1));  
run;
```

Simple Method: 2

⚠ Level_5_Text	Ⓜ len ↓	⚠ cmdecod	⚠ cmdecod1
ALANINE;ARGININE;CAL...	303	ALANINE;ARGININE...	DIBASIC;PROLINE;SE...
ALANINE;ARGININE;CAL...	303	ALANINE;ARGININE...	DIBASIC;PROLINE;SE...
ASCORBIC ACID;BIOTIN;...	263	ASCORBIC ACID;BI...	NOS;PYRIDOXINE;RE...
(RS)-3 METHYL-2-OXOVA...	230	(RS)-3 METHYL-2-O...	ACETATE;THREONIN...
(RS)-3 METHYL-2-OXOVA...	230	(RS)-3 METHYL-2-O...	ACETATE;THREONIN...
ANGELICA ACUTILOBA R...	201	ANGELICA ACUTIL...	ROOT
ASCORBIC ACID;BIOTIN;...	193	ASCORBIC ACID;BI...	
ARGININE;BETA-ALANIN...	184	ARGININE;BETA-AL...	

Simple Method: 3

```
data dv1;
set source.dv;
if length(Describe_Deviation)>200 then do;
  if substr(strip(Describe_Deviation),201,1)="" or
    substr(strip(Describe_Deviation),200,1)="" then do;
    dvterm=substr(strip(Describe_Deviation),1,200);
  end;
  else if substr(strip(Describe_Deviation),200,1)^="" then do;
    term1=substr(strip(Describe_Deviation),1,200);
    term2=reverse(term1);
    term3=scan(strip(term2),1,"");
    len1=200-length(term3);
    dvterm=substr(strip(term1),1,len1);
  end;
end;
else if length(Describe_Deviation)<=200 then dvterm=strip(Describe_Deviation);
*DVTERM2;
if length(Describe_Deviation)>200 then length=length(strip(Describe_Deviation))-200;
if length(Describe_Deviation)>200 and (substr(strip(Describe_Deviation),201,1)=""
or substr(strip(Describe_Deviation),200,1)="" ) then do;
  dvterm2=strip(substr(strip(Describe_Deviation),201,length));
end;
else if length(Describe_Deviation)>200 and substr(strip(Describe_Deviation),200,1)^="" then do;
  term4=reverse(term3);
  dvterm2=strip(term4)||strip(substr(strip(Describe_Deviation),201,length));
end;
run;
```

Simple Method: 3

⚠ dvterm	⚠ dvterm2	⚠ Describe_Deviation ↑	⊕ len
0263-303A : site a...	missing on ICF.	0263-303A : site address is present ...	213
0351-301A: D15 v...		0351-301A: D15 visit:The PK post-d...	188
0351-301A: Rand...	arm.	0351-301A: Randomization visit:The...	204
0351-301A: The H...		0351-301A: The HB for randomizati...	100
0351-301A:The sit...		0351-301A:The site is performing ax...	176
0351-301A:The sit...		0351-301A:The site is performing ax...	180

Method 4 a macro with Do Loop:

```
%split(indsn= <input dataset name>,  
outdsn= <output dataset name>,  
invar= <input variable with large text>);
```

Method 4 a macro with Do Loop:

Step 1: Find out the maximum length of the variable.

```
proc sql feedback noprint;  
    select distinct max(length(strip(&invar))) into :maxlen from &indsn;  
    %let maxlen=%trim(%left(&maxlen.));  
    %put maximum length of deter is &maxlen.;  
quit;
```

Method 4 a macro with Do Loop:

Step 2: Calculate possible number of variables to split.

```
data a;  
  set &indsn;  
  &invar = compbl(compress(&invar, 'kw'));  
  dleng=length(strip(&invar));  
  leng=floor(dleng/200);  
run;  
proc sql noprint feedback;  
select max(leng)+1 into : obs from a;  
quit;  
%put &obs;
```

Method 4 a macro with Do Loop:

Step 3: write a do loop to repeat splitting until last occurrence.

```
data &outdsn;  
set &indsn;  
length devtext $7000. ;  
dleng=length(strip(&invar));  
devtext=strip(&invar);  
%do i= 1 %to &obs;  
    length DEEVDC&i $200;  
    pos=findc(devtext, ' ', -201);  
    DEEVDC&i=strip(substr(devtext, 1, pos));  
    devtext=strip(substr(devtext,pos+1));  
%end;  
run;
```

Method 4 a macro with Do Loop:

⚠ var1	⚠ DEEVDC1	⚠ DEEVDC2	⚠ DEEVDC3
A giant contai...	A giant container ship ...	they hoped the ship co...	side.

Conclusion:

- 1) These methods will identify length of the text and split into multiple variables and that can be moved into SUPP datasets as per CDISC standards.
- 2) Above 4 methods are easy to use in any folder and any dataset including Tables and Listings.
- 3) First 3 methods need manual programming effort but method 4 will work for any text and creates n number of variables without programming effort.
- 4) The macro can identify and/or split into additional variables in a readable way complying CDISC standards
- 5) The scope of this macro is for all datasets in a specified library

Questions?

Savithri Jajam

Covance, Inc.

4000 Centergreen Way # 300

Cary, NC 27513

Savithri Jajam: savithri.jajam@covance.com

Vamshi Matta: vamshi.matta@covance.com

Lavanya Peddibhotla: lavanya.peddibhotla@covance.com



Covance by Labcorp is a leading global life sciences company, which provides contract research services to the drug, medical device and diagnostics, crop protection and chemical industries. COVANCE is a registered trademark and the marketing name for Covance Inc. and its subsidiaries around the world.

www.covance.com