SAS ユーザー総会 2017

レガシーデータ変換等における トランスコーディングの問題に関する 考察と提案

○山崎文寛[1]、高浪洋平[1]

([1]武田薬品工業株式会社)

Considerations and Proposals on Transcoding issues in Legacy Data Conversion

Fumihiro Yamasaki, Yohei Takanami
Takeda Pharmaceutical Company

要旨

新医薬品の製造販売の承認申請時電子データ提出におけるレガシーデータ変換時のトランスコーディングに関する問題を整理し，医薬品医療機器総合機構 (PMDA) の規制要件に合致しないデータを検出・変換する方法を提案する

キーワード: トランスコーディング, セッションエンコーディング, 符号化方式 (Encoding), 文字セット (Character Set), ASCII, Shift-JIS, Wlatin1, UTF-8, 電子データ提出，統合解析，レガシーデータ変換

本日の内容

1. 発表の背景

2. 文字セットと符号化方式

- PMDAの規制要件
- 文字セットASCII
- 文字セットUnicodeと符号化方式UTF-8
- 非ASCII文字及び非印字可能文字の検出と変換
- トランスコーディングの必要性

3. 非ASCII文字及び非印字可能文字の検出・変換ツールの開発事例

- テストデータ
- ツールの構成
- ツールの実行

4. まとめ

本日の内容

1. 発表の背景

2. 文字セットと符号化方式
   - PMDAの規制要件
   - 文字セットASCII
   - 文字セットUnicodeと符号化方式UTF-8
   - 非ASCII文字及び非印字可能文字の検出と変換
   - トランスコーディングの必要性

3. 非ASCII文字及び非印字可能文字の検出・変換ツールの開発事例
   - テストデータ
   - ツールの構成
   - ツールの実行

4. まとめ

発表の背景

本邦では，2016年10月1日より承認申請時の電子データ提出が開始され，各製薬企業はCDISC標準に準拠した電子データの提出準備を進めている．その中で，過去に国内外で実施した臨床試験について電子データを提出する場合は，規制要件に合致するようにレガシーデータ変換を行っている

レガシーデータ変換に用いるSASシステムの符号化方式* が，レガシーデータ作成時のそれと異なる場合，SASデータセット読込時にエラーが発生し得る．これは符号化方式ごとに，表現可能な文字や1文字あたりのバイト数が異なるためである．また，エラーが発生しなくとも，規制要件に合致しない文字がレガシーデータに含まれる可能性がある

本稿では，レガシーデータ変換業務等を効率的に実施することを目的として，規制要件に合致しないデータをエラーなく適切に検出・変換する方法を検討することとした

*符号化方式：情報をコンピュータ上で処理するためにデジタルデータ化すること

本日の内容

1. 発表の背景

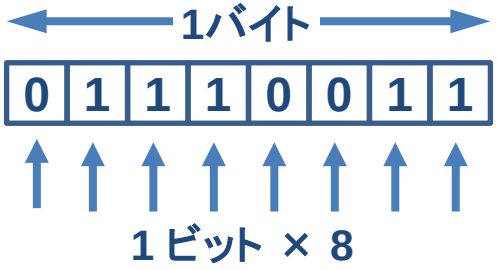
2. 文字セットと符号化方式
   - PMDAの規制要件
   - 文字セットASCII
   - 文字セットUnicodeと符号化方式UTF-8
   - 非ASCII文字及び非印字可能文字の検出と変換
   - トランスコーディングの必要性

3. 非ASCII文字及び非印字可能文字の検出・変換ツールの開発事例
   - テストデータ
   - ツールの構成
   - ツールの実行

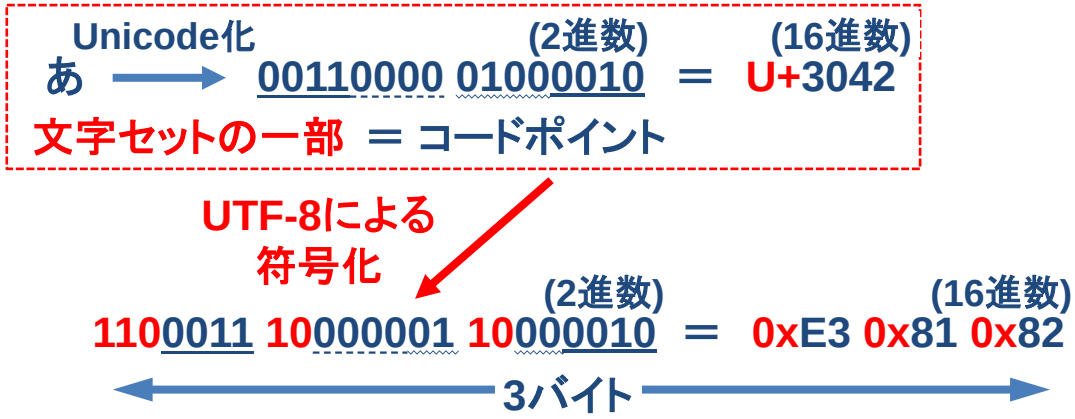
4. まとめ

文字セットと符号化方式

- ビットとバイト



- 文字セットと符号化方式



- SASデータセット使用時に主に使用される文字セットと符号化方式の関係

文字セット	符号化方式	SASで指定する名称 (短縮名)	SAS上の 最大バイト数
JISX0208	Shift-JIS	shift-jis (sjis)	2
	EUC	euc-jp (jeuc)	4
Unicode	UTF-8	utf-8 (utf8)	4
	UTF-16BE	utf-16b	2
	UTF-16LE	utf-16l	2
Extended ASCII	Code Page 1252 (Windows Latin-1)	wlatin1 (wlt1)	1

PMDAの規制要件

- 提出するデータセットに関する要件
  - ✓ 日本語項目がドメインに含まれる場合の対応方法

※有害事象の医師記載名，コメント等が日本語

- 承認申請時の電子データ提出等に関する技術的ガイドについて (平成28年8月24日付け薬機次発第0824001号 独立行政法人医薬品医療機器総合機構次世代審査等推進室長通知により改正)

✓ 申請電子データ提出時に用いられるバリデーションルール

- PMDA Study Data Validation Rules (V1.0, 2015/11/18)
- 非ASCII文字及び非印字可能文字 (ASCIIコードの32から126以外の値) が変数の値に含まれる場合はWarning (SDTM Rules V1.0, RULE ID: SD1029)

✓ 申請時に提示すべき文字セット及び符号化方式の情報の具体例

- PMDA 申請電子データに関するFAQ (Q4-19) から抜粋

【文字セット】JISX0208	【符号化方式】Shift-JIS
【文字セット】JISX0208	【符号化方式】EUC-JP
【文字セット】UNICODE (USC-2)	【符号化方式】UTF-8

文字セットASCII

- 特徴
  - ✓ 7 ビットで表現可能な非負整数の0～127に対して，アルファベット，数字等を割当てたもの
  - ✓ 0 ～ 31, 127番目：制御文字 (改行, 後退, タブ等), 32 ～ 126番目：印字可能文字 (32番目：半角スペース, 48～57：数字, 65～90, 97～122：大小アルファベット, それ以外：各種記号)

- ASCII一覧 (32 ～ 126番目)

32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51
(space)	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/	0	1	2	3
52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
4	5	6	7	8	9	:	;	<	=	>	?	@	A	B	C	D	E	F	G
72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91
H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	[
92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111
¥	]	^	_	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
112	113	114	115	116	117	118	119	120	121	122	123	124	125	126					
p	q	r	s	t	u	v	w	x	y	z	{	|	}	~					

- 英語データセットを処理する際の注意点
  - ✓ ASCIIの128～256番目を追加した拡張ASCII＋符号化方式ISO 8859が存在
  - ✓ 英語圏で作成されたSASデータセットは，**ISO 8859-1 (Windows Latin-1)** 等の符号化方式が使用されている可能性がある ⇒ **PMDAへの電子データ提出時に要注意！**

文字セットUnicodeと符号化方式UTF-8

- デフォルトのセッションエンコーディング

Locale＼OS	Windows	UNIX	z/OS
English_US (en_US)	WLatin1 (wlt1)	Latin1	open_ed-1047
Japanese_Japan (jp_JP)	Shift-JIS (sjis)	Shift-JIS, EUC-JP等	ibm-939

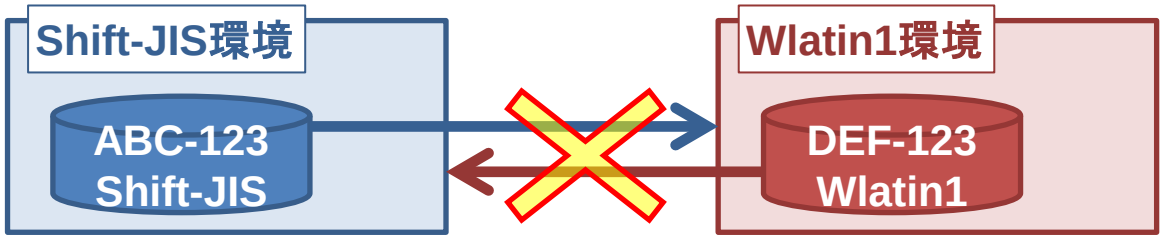
- Unicodeとは？
  - ✓ 世界中のほぼすべての言語における文字に，ユニークなコードを割り当てた文字セット
  - ✓ Unicodeの最初の128コードはASCIIと，最初の256コードはISO 8859-1と同一
  - ✓ 日本語文字セットのJISX0208等，一般的に使用される多くの文字セットと互換性あり

- 各符号化方式の比較

符号化方式		UTF-8	Shift-JIS	Wlatin1
文字セット		Unicode	JISX0208	Extended ASCII
日本語表示		○	○	×
バイト数	ASCII部分 (英数字, 記号)	1バイト	1バイト	1バイト
	ASCII以外	2～4バイト	1～2バイト	1バイト
	平仮名, 片仮名	3バイト	2バイト	×
	漢字	3バイト(一部4バイト)	2バイト	×
	拡張ASCIIの128番目以降 (ü, Ä, ®など)	2バイト	×	1バイト

文字セットUnicodeと符号化方式UTF-8 (続き)

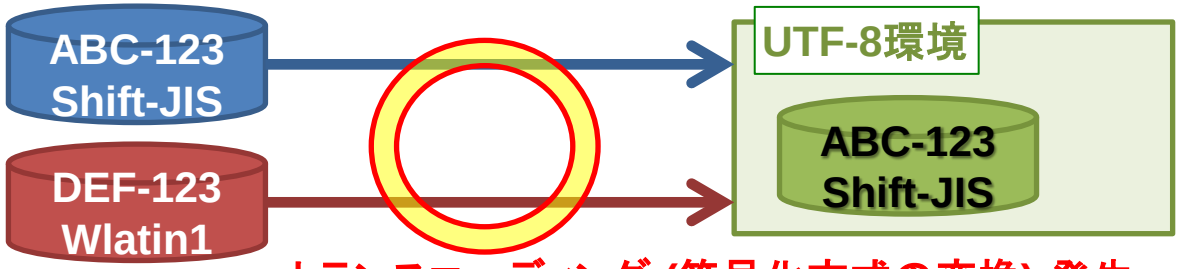
- レガシーデータにおける符号化方式



互いの環境で読み込めない文字が存在

例）Wlatin1環境で日本語データは読み込めない

- Unicode環境を使用



トランスコーディング (符号化方式の変換) 発生
⇒ 1文字あたりのバイト数が変化 ⇒ 変数値に対して変数の長さが不足

✓ LIBNAMEステートメントのオプションを用いたエラー回避方法
⇒ SASデータセット読込時に変数の長さを一律で変更

CVPBYTES=数字	トランスコーディング時に各変数で拡張するバイト数を指定 例)「CVPBYTES=5」を指定⇒変数の長さが5の場合は10に，10の場合は15
CVPMULTIPLIER=数字	トランスコーディング時に各変数で拡張する乗数値を指定 例)「CVPMULTIPLIER=1.5」を指定⇒変数の長さが10の場合は15

非ASCII文字及び非印字可能文字の検出と変換

- Srivastava (2017) の方法を参考にした検出マクロプログラム
  - ✓ 検出結果をExcelに出力　　✓ 主にUTF-8環境での実行を想定
  - ✓ BYTE関数を使用してASCII文字のリストを作成　⇒　変数値と照合 (例：byte(97) ⇒ 'a')

%DetectNonPriASCII (DataFolder=, KeyVariables=, CVPMult=) ;	
DataFolder	検出対象のSASデータセットを格納したフルパス (必須)
KeyVariables	検出結果とともに表示させる被験者番号等のキー変数 (オプション) (複数指定の場合は半角スペース区切り)
CVPMult	各変数の長さを一律拡張するための乗数値 (オプション) ※LIBNAMEステートメントのCVPMULTIPLIERオプションに指定される値

- 実行結果

	A	B	C	D	E	F	G	H	I
1	OBS	STUDYID	USUBJID	OBS	MEMNAME	NAME	ORGVALUE	DetectedString	CandidateValue
2	1	ABC-123	ABC-123-1001-001	2	ADAE	AETERM	NEUTROPENIA (0.86 MILLE/MM³)	³	
3	2	ABC-123	ABC-123-1001-002	3	ADAE	AETERM	BAKER´S CYST	´	
4	3	ABC-123	ABC-123-1002-001	5	ADAE	AETERM	ANÉMIA	É	
5	4	ABC-123	ABC-123-1003-001	7	ADAE	AETERM	INFECTION IN OUTER EAR (PLUS DRANÆGE	Æ	
6	5	ABC-123	ABC-123-1003-002	9	ADAE	AETERM	DERMO-HYPODERMITIS ON LEFT WRIST		
7	6	ABC-123	ABC-123-1003-003	11	ADAE	AETERM	ASTHÉNIA	É	
8	7	ABC-123	ABC-123-1003-004	12	ADAE	AETERM	ÖDEMA LOWER LIMBS	Ö	
9	8	ABC-123	ABC-123-1003-004	13	ADAE	AETERM	ÖDEMA RIGHT FOOT	Ö	

KeyVariables

オブザベーション番号

データセット名

変数名

検出結果が含まれる変数値の全体

検出値

非ASCII文字及び非印字可能文字の検出と変換 (続き)

- 【参考】ASCII文字のリスト作成用プログラムと実行結果

```
data ListOfASCII;
   length ASCII_List $200.;
   ASCII_List='';
   do i=32 to 126;
      ASCII_List=trim(ASCII_List)||byte(i);
   end;
run;
```

	ASCII_List
1	!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[¥]^_`abcdefghijklmnopqrstuvwxyz{|}~

- 【参考】非ASCII及び非印字可能文字の検出

```
data ＜検出結果データセット＞;
   set ＜検出対象データセット＞;
   DetectedString=kcompress(＜検出対象の変数＞,ASCII_List);
run;
```

※事前に検出対象データセットの全レコードにASCII文字のリスト変数をマージしておく

非ASCII文字及び非印字可能文字の検出と変換 (続き)

- 変換マクロプログラム
  - ✓ 検出結果 (Excel) に変換したい値を入力し，その値で変数値ごと変換
  - ✓ 符号化方式の変換，変数長の変数値の最大長への変換

%ReplaceNonAscii (InDataFolder= ,OutDataFolder= ,InExcelFile= ,InExcelSheet= , OutEncoding=,CVPMult=) ;	
InDataFolder	変換対象のSASデータセットを格納したフルパス (必須)
OutDataFolder	変換後のSASデータセットを出力するフルパス (必須)
InExcelFile	変換候補の値を追記したファイルのフルパス (必須)
InExcelSheet	変換候補の値を追記したファイルにおけるシート名 (必須)
OutEncoding	変換後の符号化方式 (デフォルトはUTF-8)
CVPMult	各変数の長さを一律拡張するための乗数値 (オプション)

- 変換結果

Excelファイル

Candidate Value (I列) に変換したい値を入力

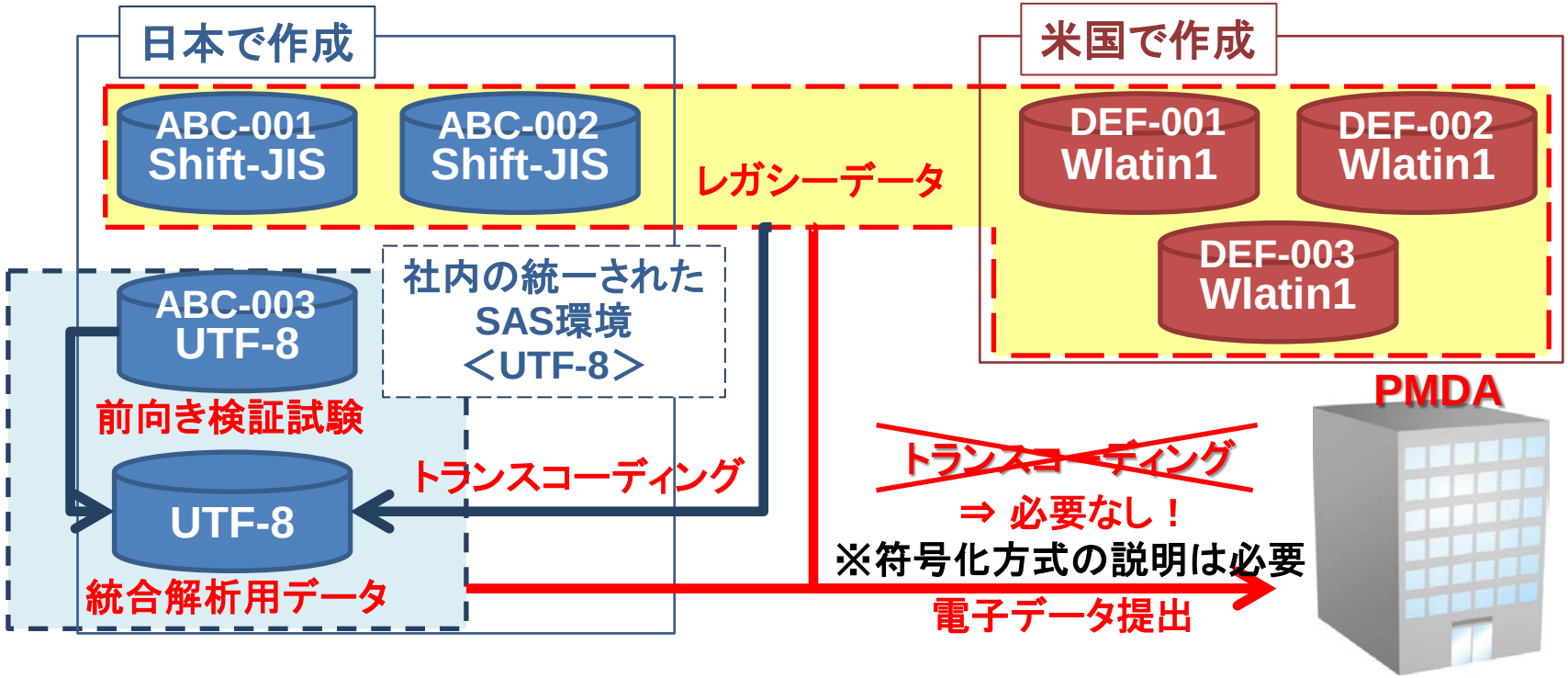
A	B	E	F	G	H	
OBS	USUBJID	OBS	MEMNAME	NAME	ORGVALUE	DetectedString	CandidateValue
1	ABC-123-1001-001	Wound infection	1	ADAE_SJIS	AETERM	左足親指の傷感染	左足親指の傷感染	JAPANESE TEXT IN SOURCE DATABASE
2	ABC-123-1001-001	Neutropenia	2	ADAE_SJIS	AETERM	好中球減少 (0.86 MILLE/MM^3)	好中球減少	JAPANESE TEXT IN SOURCE DATABASE
3	ABC-123-1001-002	Synovial cyst	3	ADAE_SJIS	AETERM	ベーカー嚢胞	ベーカー嚢胞	JAPANESE TEXT IN SOURCE DATABASE
4	ABC-123-1002-001	Paraes						SE TEXT IN SOURCE DATABASE
5	ABC-123-1002-001	Anaem						SE TEXT IN SOURCE DATABASE

SASデータセット

	STUDYID	USUBJID	AESEQ	AETERM	AEDECOD
1	A		1	JAPANESE TEXT IN SOURCE DATABASE	Wound infection
2	A		2	JAPANESE TEXT IN SOURCE DATABASE	Neutropenia
3	ABC-123	ABC-123-1001-002	1	JAPANESE TEXT IN SOURCE DATABASE	Synovial cyst
4	ABC-123	ABC-123-1002-001	1	JAPANESE TEXT IN SOURCE DATABASE	Paraesthesia
5	ABC-123	ABC-123-1002-001	2	JAPANESE TEXT IN SOURCE DATABASE	Anaemia

変換

トランスコーディングの必要性



- 意図しないトランスコーディングは，処理速度の低下，文字化け及び文字切れの原因となるため，UTF-8等の符号化方式に変換・統一しておくことも選択肢の一つ
  - トランスコーディングマクロプログラム

%Transcoding2UTF8 (InDataFolder=, OutDataFolder=, Outencoding=) ;	
InDataFolder	変換対象のSASデータセットを格納したフルパス (必須)
OutDataFolder	変換後のSASデータセットを出力するフルパス (必須)
Outencoding	変換後の符号化方式 (必須)

本日の内容

1. 発表の背景

2. 文字セットと符号化方式
   - PMDAの規制要件
   - 文字セットASCII
   - 文字セットUnicodeと符号化方式UTF-8
   - 非ASCII文字及び非印字可能文字の検出と変換
   - トランスコーディングの必要性

3. **非ASCII文字及び非印字可能文字の検出・変換ツールの開発事例**
   - テストデータ
   - ツールの構成
   - ツールの実行

4. まとめ

検出・変換ツールの開発事例

- 非ASCII文字及び非印字可能文字 (ASCIIの32から126番目以外の文字) の検出，並びにそれらのASCII文字への変換をサポートするツール
- SAS及びHTMLアプリケーション (HTA) を使用
  - ✓ Windows上で動作可能なアプリケーション
  - ✓ HTML及びJava Script等の機能を利用して手軽なGUI画面作成が可能
  - ✓ 基本的な機能及びSASとの連携に関する詳細は舟尾・高浪 (2008)及び高浪 (2012) を参照

- テストデータ
  - ✓ 米国で実施された仮想の臨床試験で得られた有害事象に関する解析用データセット

Wlatin1

STUDYID	USUBJID	AESEQ	AETERM	・・・
ABC-123	ABC-123-1001-001	1	WOUND INFECTION LEFT BIG TOE	・・・
ABC-123	ABC-123-1001-001	2	NEUTROPENIA (0.86 MILLE/MM^3)	・・・
ABC-123	ABC-123-1001-002	1	BAKER´S CYST	・・・
・・・	・・・	・・・	・・・	・・・

  - ✓ 日本で実施された仮想の臨床試験で得られた有害事象に関する解析用データセット

Shift-JIS

STUDYID	USUBJID	AESEQ	AETERM	・・・
ABC-123	ABC-123-1001-001	1	左足親指の傷感染	・・・
ABC-123	ABC-123-1001-001	2	好中球減少 (0.86 MILLE/MM^3)	・・・
ABC-123	ABC-123-1001-002	1	ベーカー嚢胞	・・・
・・・	・・・	・・・	・・・	・・・

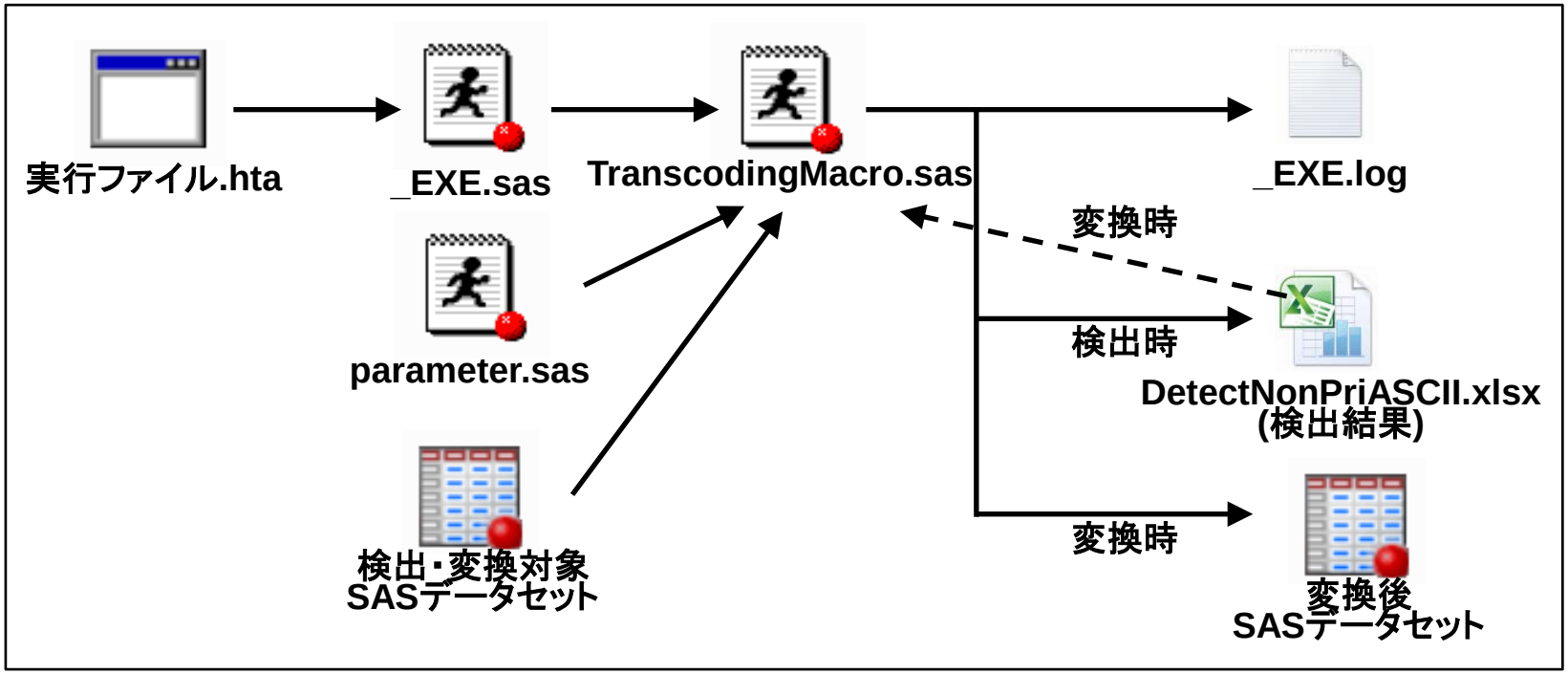
ツールの構成

➢ 各フォルダ及びファイルの構成

実行ファイル .hta	:本ファイルをダブルクリックするとツールが起動	
files フォルダ	_EXE.log	:実行時のログ ※実行時に自動作成される
	_EXE.sas	:実行用の一時ファイル (各SASプログラムを%INCLUDEステートメントで実行) ※実行時に自動作成される
	parameter.sas	:実行用の一時ファイル (画面から与えた情報をSASマクロ変数化) ※実行時に自動作成される
	TranscodingMacro.sas	:非ASCII文字及び非印字可能文字の検出・変換，符号化方式の変換を行うSASマクロ．本稿で作成した以下のSASマクロプログラムを使用 ・DetectNonPriASCII ・ReplaceNonAscii ・Transcoding2UTF8
	DetectNonPriASCII.xlsx	:非ASCII文字及び非印字可能文字の検出結果ファイル，また変換候補文字を入力するファイル ※実行時に自動作成される
INDATA フォルダ	:検出・変換対象のSASデータセットを格納するフォルダ ※任意の場所を指定可能	
OUTDATA フォルダ	:検出・変換結果のSASデータセットが格納されるフォルダ ※任意の場所を指定可能	

ツールの構成

- ツールの処理の流れ

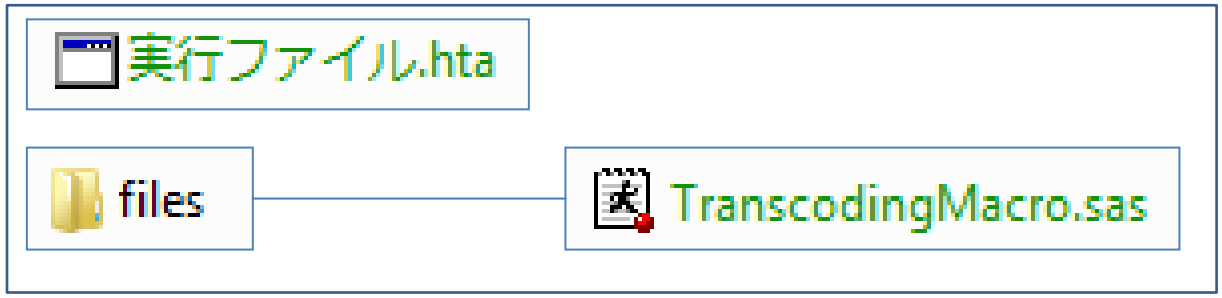


ツールの実行

- ツール実行前に変更すること
  - ✓ 実行ファイル.htaをテキストエディタで開き，SASの実行可能ファイル (.exe)及び環境設定ファイル(.cfg)のパスをSASのインストール環境に合わせて変更
  - ✓ 環境設定ファイルはパスに「u8」が含まれるパスを指定 (SAS Unicodeサポートを使用)

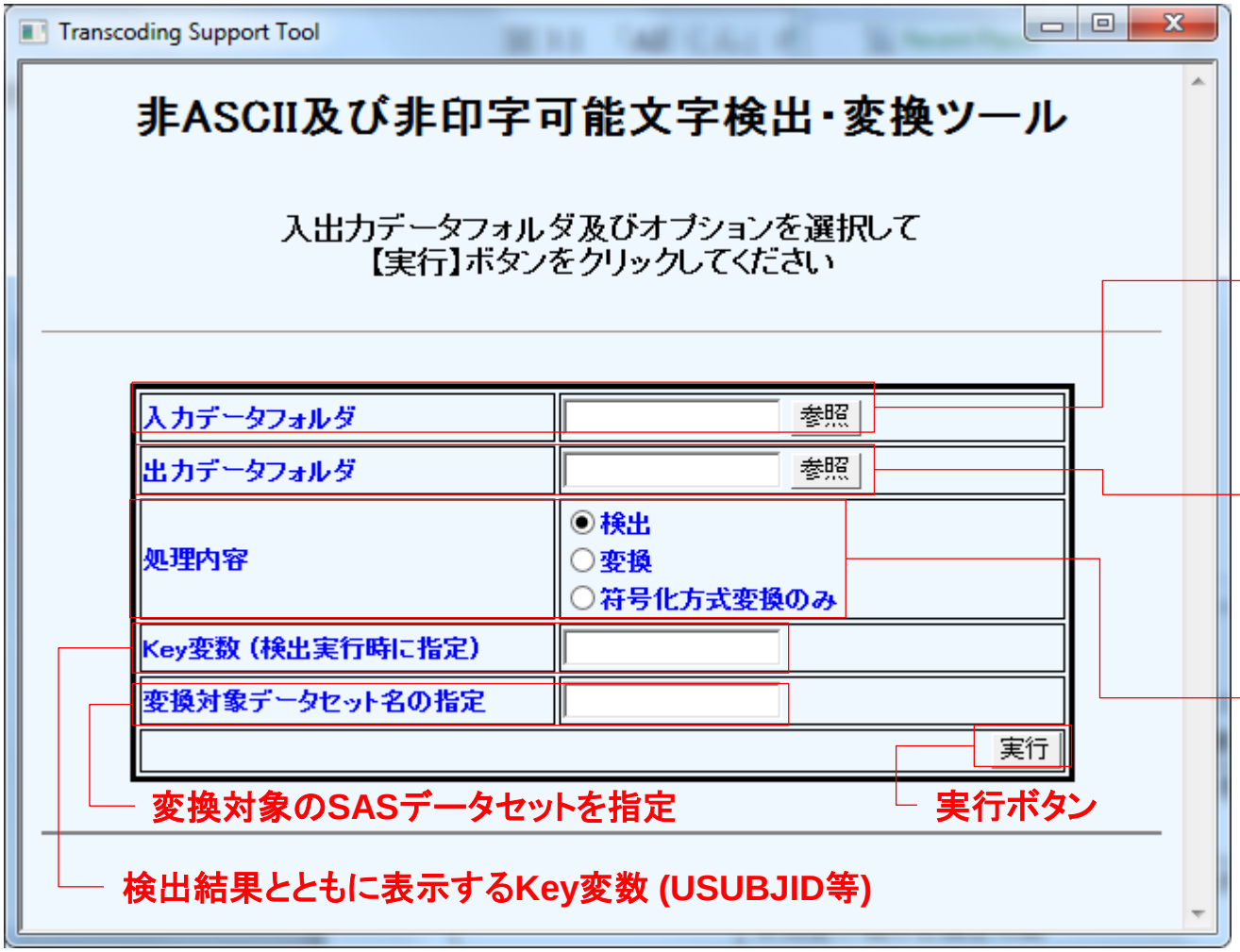
```
// _EXE.sasの作成
  MyText = MyScript.CreateTextFile("./files/_EXE.sas", true) ;
  MyText.WriteLine("%inc '" + MyPath + "¥¥files¥¥parameter.sas' ;") ;
  MyText.WriteLine("%inc '" + MyPath + "¥¥files¥¥TranscodingMacro.sas' ;") ;
  MyText.Close() ;
// Shell関連の操作を提供するオブジェクトを取得
  MyShell = new ActiveXObject("WScript.Shell") ;
// SASの実行
MyShell.Exec('"C:/Program Files/SAS_9.4/SASFoundation/9.4/sas.exe"
              -CONFIG "C:/Program Files/SAS_9.4/SASFoundation/9.4/nls/u8/sasv9.cfg"
              -sysin "./files/_EXE.sas" -log "./files/_EXE.log" -nosplash -icon') ;
```

  - ✓ ツール一式を任意の場所に設置
    例) C:¥Testの直下に，実行ファイル.hta及びfilesフォルダを設置



ツールの実行

- ツールの画面及び指定可能なパラメータ



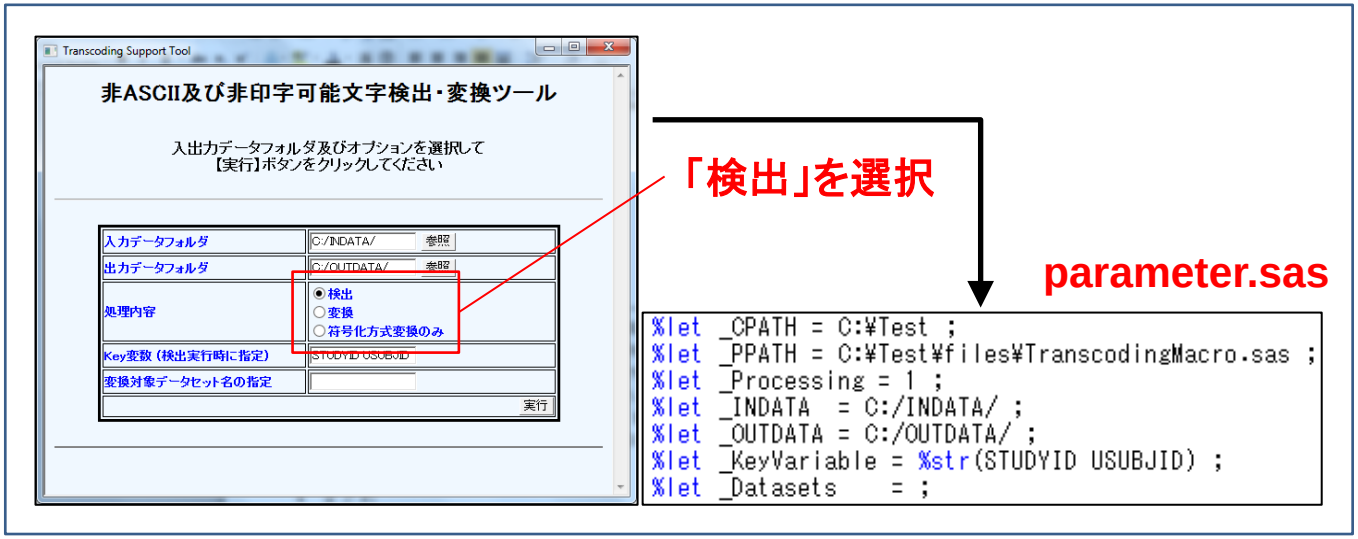
検出・変換対象のSASデータセットの格納フォルダを選択 (「参照」ボタンをクリックして選択)

変換結果のSASデータセットの格納フォルダを選択 (「参照」ボタンをクリックして選択)

処理内容を選択．「変換」及び「符号化方式変換のみ」は，本ツールではUTF-8に変換される

ツールの実行

- 非ASCII文字及び非印字可能文字の検出
  - ✓画面からのパラメータの入力　⇒ SASプログラムの生成



✓検出結果 (DetectNonPriASCII.xlsx)

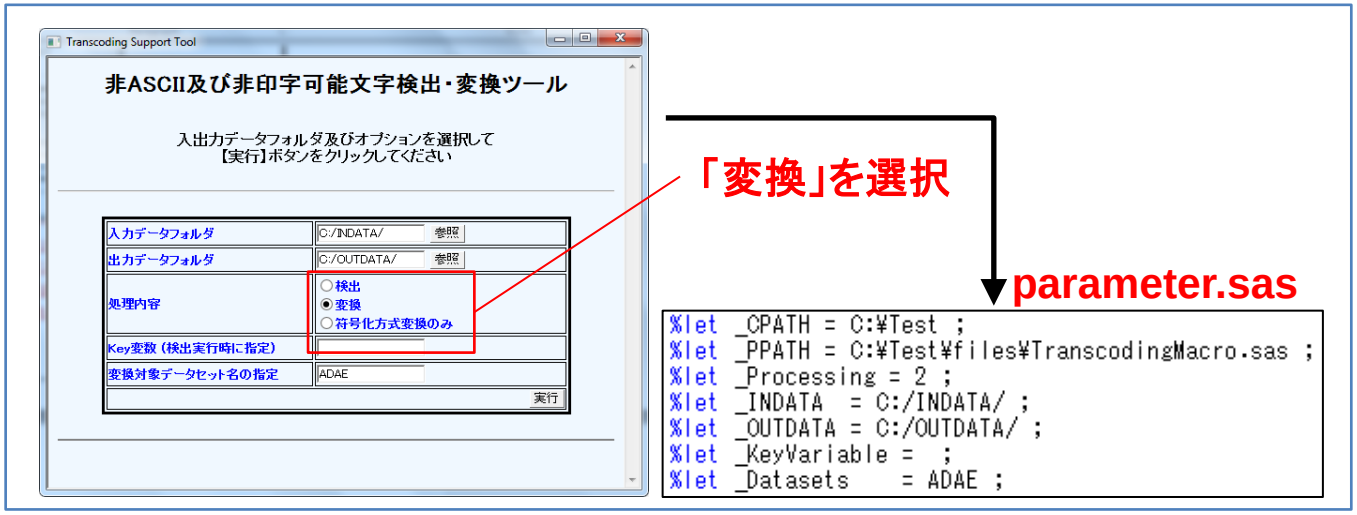
※日本で実施された仮想の臨床試験で得られた
有害事象に関する解析用データセットを使用

	A	B	C	D	E	F	G	H	I
1	OBS	STUDYID	USUBJID	OBS	MEMNAME	NAME	ORGVALUE	DetectedString	CandidateValue
2	1	ABC-123	ABC-123-1001-001	1	ADAE	AETERM	左足親指の傷感染	左足親指の傷感染	
3	2	ABC-123	ABC-123-1001-001	2	ADAE	AETERM	好中球減少 (0.86 MILLE/MM^3)	好中球減少	
4	3	ABC-123	ABC-123-1001-002	3	ADAE	AETERM	ベーカー嚢胞	ベーカー嚢胞	
5	4	ABC-123	ABC-123-1002-001	4	ADAE	AETERM	下肢のうずき	下肢のうずき	
6	5	ABC-123	ABC-123-1002-001	5	ADAE	AETERM	貧血	貧血	
7	6	ABC-123	ABC-123-1003-001	6	ADAE	AETERM	ステロイド誘発障害	ステロイド誘発障害	
8	7	ABC-123	ABC-123-1003-001	7	ADAE	AETERM	外耳感染症	外耳感染症	
9	8	ABC-123	ABC-123-1003-002	8	ADAE	AETERM	右手の痙攣	右手の痙攣	
10	9	ABC-123	ABC-123-1003-002	9	ADAE	AETERM	左手首の皮膚皮下組織炎	左手首の皮膚皮下組織炎	
11	10	ABC-123	ABC-123-1003-003	10	ADAE	AETERM	便秘	便秘	
12	11	ABC-123	ABC-123-1003-003	11	ADAE	AETERM	衰弱	衰弱	
13	12	ABC-123	ABC-123-1003-004	12	ADAE	AETERM	下肢の浮腫	下肢の浮腫	
14	13	ABC-123	ABC-123-1003-004	13	ADAE	AETERM	右足の浮腫	右足の浮腫	

変換が必要な場合，
変換したい値を入力
例) JAPANESE TEXT IN SOURCE DATABASE

ツールの実行

- 非ASCII文字及び非印字可能文字の変換
  - ✓画面からのパラメータの入力　⇒ SASプログラムの生成



✓変換結果

	STUDYID	USUBJID	AESEQ	AETERM	AEDECOD
1	ABC-123	ABC-123-1001-001	1	JAPANESE TEXT IN SOURCE DATABASE	Wound infection
2	ABC-123	ABC-123-1001-001	2	JAPANESE TEXT IN SOURCE DATABASE	Neutropenia
3	ABC-123	ABC-123-1001-002	1	JAPANESE TEXT IN SOURCE DATABASE	Synovial cyst
4	ABC-123	ABC-123-1002-001	1	JAPANESE TEXT IN SOURCE DATABASE	Paraesthesia
5	ABC-123	ABC-123-1002-001	2	JAPANESE TEXT IN SOURCE DATABASE	Anaemia

ツールの実行

- 処理として「変換」又は「符号化方式の変換のみ」を選択した場合，以下のように符号化方式が変換される

データセット名	INDATA.ADAE
メンバータイプ	DATA
エンジン	V9
作成日時	2017/06/20 09:22:34
更新日時	2017/06/20 09:22:34
保護	
データセットタイプ	
ラベル	
データ表現	WINDOWS_32
エンコード	shift-jis Japanese (SJIS)

データセット名	OUTDATA.ADAE
メンバータイプ	DATA
エンジン	V9
作成日時	2017/06/20 13:52:10
更新日時	2017/06/20 13:52:10
保護	
データセットタイプ	
ラベル	
データ表現	WINDOWS_32
エンコード	utf-8 Unicode (UTF-8)

- 変数の長さのデータの最大長への調整も同時に実施される

#	変数	タイプ	長さ	ラベル
1	STUDYID	文字	7	Study Identifier
2	USUBJID	文字	16	Unique Subject Identifier
3	AESEQ	数値	8	Sequence Number
4	AETERM	文字	100	Reported Term for the Adverse Event
5	AEDECOD	文字	25	Dictionary-Derived Term
6	AEBODSYS	文字	52	Body System or Organ Class
7	TRTEMFL	文字	1	Treatment Emergent Analysis Flag
8	AESTDTC	文字	10	Start Date/Time of Adverse Event
9	AEENDTC	文字	10	End Date/Time of Adverse Event
10	AESER	文字	1	Serious Event
11	AESEV	文字	8	Severity/Intensity
12	AEREL	文字	11	Causality

#	変数	タイプ	長さ	ラベル
1	STUDYID	文字	7	Study Identifier
2	USUBJID	文字	16	Unique Subject Identifier
3	AESEQ	数値	8	Sequence Number
4	AETERM	文字	32	Reported Term for the Adverse Event
5	AEDECOD	文字	25	Dictionary-Derived Term
6	AEBODSYS	文字	52	Body System or Organ Class
7	TRTEMFL	文字	1	Treatment Emergent Analysis Flag
8	AESTDTC	文字	10	Start Date/Time of Adverse Event
9	AEENDTC	文字	10	End Date/Time of Adverse Event
10	AESER	文字	1	Serious Event
11	AESEV	文字	8	Severity/Intensity
12	AEREL	文字	11	Causality

本日の内容

1. 発表の背景

2. 文字セットと符号化方式
   - PMDAの規制要件
   - 文字セットASCII
   - 文字セットUnicodeと符号化方式UTF-8
   - 非ASCII文字及び非印字可能文字の検出と変換
   - トランスコーディングの必要性

3. 非ASCII文字及び非印字可能文字の検出・変換ツールの開発事例
   - テストデータ
   - ツールの構成
   - ツールの実行

4. まとめ

まとめ

文字セットと符号化方式

- 文字セットと符号化方式は，SASデータセットの作成に用いたSAS環境に依存
- 統合解析用データセットの作成時等にトランスコーディングが発生する可能性あり
- PMDAは，英数字等のASCIIで規定された文字セットのみで構成されたデータを要求

非ASCII文字及び非印字可能文字の検出・変換

- HTAを用いた，ラジオボタンや各種入力フィールド等から構成されるユーザーフレンドリーな画面からSASプログラムを実行する方法を適用した，非ASCII文字及び非印字可能文字の検出・変換ツールの開発事例を紹介

海外で作成された電子データをPMDAに提出する際の注意点

- PMDAの規制要件に合致しない非ASCII文字及び非印字可能文字が含まれる可能性があることに注意

参考文献

- Hui Song et al. (2016). The Impact of Change from wlatin1 to UTF-8 Encoding in SAS Environment; PharmaSUG 2016; Paper BB15
- Jing Gao (2015). Multilingual data support in Dataset-XML with SAS® Clinical Data Integration; PharmaSUG China 2015; Paper 25
- Donna Dutton (2015). Data Encoding: All Characters for All Countries; PhUSE2015; Paper DH03
- Sridhar R Dodlapati et al. (2010). Non Printable & Special Characters: Problems and how to overcome them; NESUG 2010
- Abhinav Srivastva (2017). Reporting Non-Printable and Special Characters for Review in Excel; PharmaSUG2017; Paper BB04
- Manfred Kiefer (2012). SAS® Encoding: Understanding the Details; SAS Institute Inc.
- SAS Institute Inc. SAS 9.4 National Language Support (NLS): Reference Guide, Fifth Edition; Available at http://support.sas.com/documentation/onlinedoc/nls/
- SAS Institute Inc. Changing language during a SAS session
- https://support.sas.com/resources/papers/LocaleSwitching.pdf [Accessed 7 July 2017]
- Business Rules ; Available at https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm2005545.htm [Accessed 7 July 2017]
- PMDA次世代審査・相談体制について (申請電子データ提出); Available at https://www.pmda.go.jp/review-services/drug-reviews/about-reviews/p-drugs/0003.html [Accessed 7 July 2017]
- 高浪洋平 (2012). SASとHTMLアプリケーションによるCDISC ADaM形式の解析用データセットを用いた有害事象の解析帳票・グラフ簡易作成ツールの開発事例; SASユーザー総会2012
- 舟尾暢男, 高浪洋平 (2008). HTMLアプリケーションを用いた簡易SASツールの開発事例 - 臨床試験における症例数設計用ツールの構築 -; SASユーザー総会2008
- 高浪洋平, 舟尾暢男 (2015). 改訂版 統計解析ソフト「SAS」; カットシステム