WITH STANDARDS – UNLOCK THE POWER OF DATA

cdisc

2022
CHINA
INTERCHANGE
29-30 JULY | VIRTUAL EVENT

The Tool of Improving Clinical Data Quality Based on SDTM Datasets
基于SDTM数据集开发提高临床数据质量工具

Presented by Yunhui Cui, Senior SAS programmer, Program Department, GCP ClinPlus Co., Ltd.

Meet the Speaker

Yunhui Cui 崔允辉

Title: Senior SAS programmer

Organization: GCP ClinPlus Co., Ltd. 普瑞盛医药科技开发有限公司

Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*

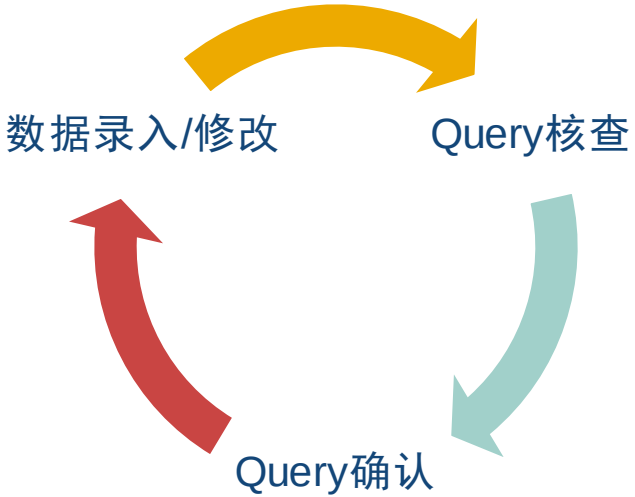
Agenda

1. 背景
2. 基于SDTM数据集开发提高临床数据质量工具
3. 实际项目运行实例
4. 总结

背景

背景

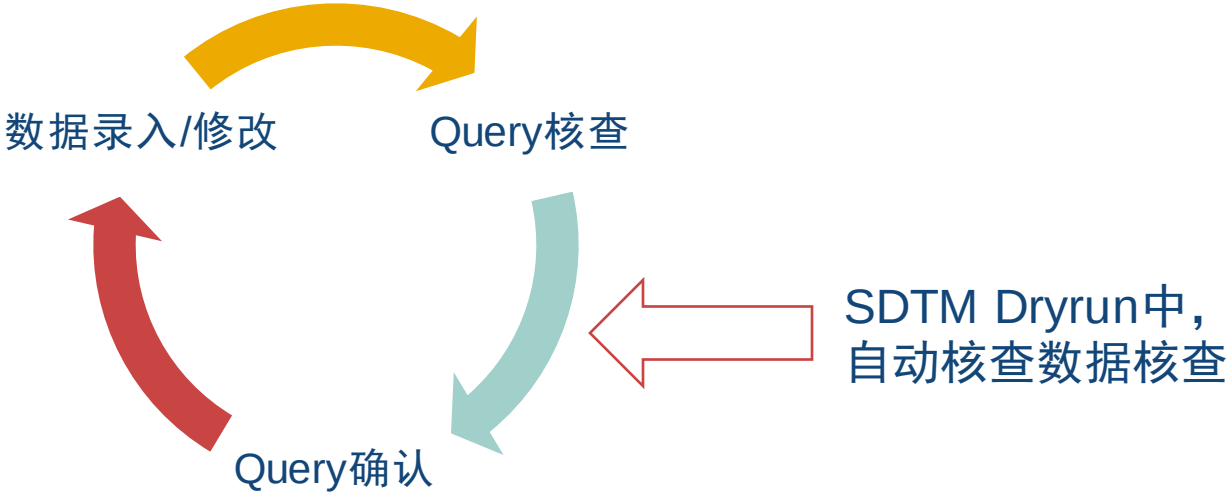
在临床过程试验中，在数据清理阶段针对数据质量核查能发现所有问题，或者对所有数据进行100%的核查，是不现实且没有必要，也不是发现错误、提高数据质量的有效方法。



背景

- 甚至在临床试验项目锁库完成后，在进行质量核查过程中，也总会发现有数据错误或者没有解决的问题出现，这时往往面临是否重新解锁但又会拖慢试验进程的艰难决定。

背景



因此，除了针对原始EDC数据按照正常数据管理流程做程序核查（数据核查和/或逻辑核查）外，基于SDTM数据集做自动化的数据或逻辑核查：

(1) 不仅能够减少锁库后出现数据问题的风险，也能够侧面印证数据管理的质量。

(2) 同时，相比基于杂乱的原始数据，在标准化的SDTM数据集基础上做有针对和重点的核查，在效率和质量上都能够达到事半功倍的效果。

为什么选择基于SDTM?

- SDTM具有稳定的数据集结构，便于实现程序自动化处理

#	Variable Name	Variable Label	Type	Format	Description
1	VISITNUM	Visit Number	Num		Clinical encounter number. Numeric version of VISIT, used for sorting.
2	VISIT	Visit Name	Char		Protocol-defined description of a clinical encounter.
3	VISITDY	Planned Study Day of Visit	Num		Planned study day of VISIT. Should be an integer.

具有标准的访视变量

为什么选择基于SDTM?

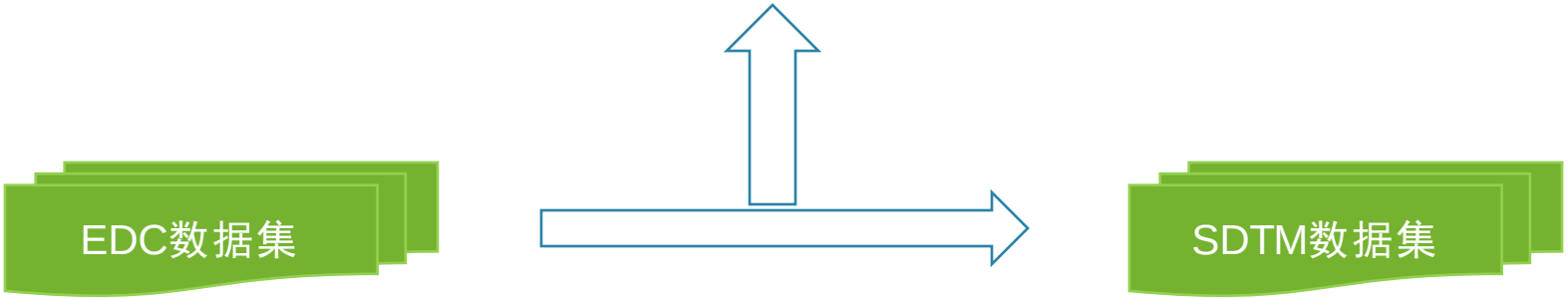
10	--DTC	Date/Time of Collection	Char	ISO 8601	Collection date and time of an observation.
11	--STDTC	Start Date/Time of Observation	Char	ISO 8601	Start date/time of an observation.
12	--ENDTC	End Date/Time of Observation	Char	ISO 8601	End date/time of the observation.
13	--DY	Study Day of Visit/Collection/Exam	Num		Actual study day of visit/collection/exam expressed in integer days relative to the sponsor-defined RFSTDTC in Demographics.
14	--STDY	Study Day of Start of Observation	Num		Actual study day of start of observation expressed in integer days relative to the sponsor-defined RFSTDTC in Demographics.
15	--ENDY	Study Day of End of Observation	Num		Actual study day of end of observation expressed in integer days relative to the sponsor-defined RFSTDTC in Demographics.

具有标准的时间变量与标准的时间格式

	Date and Time as Originally Recorded	Precision	ISO 8601 Date/Time
1	December 15, 2003 13:14:17.123	Date/time, including fractional seconds	2003-12-15T13:14:17.123
2	December 15, 2003 13:14:17	Date/time to the nearest second	2003-12-15T13:14:17
3	December 15, 2003 13:14	Unknown seconds	2003-12-15T13:14
4	December 15, 2003 13	Unknown minutes and seconds	2003-12-15T13
5	December 15, 2003	Unknown time	2003-12-15
6	December, 2003	Unknown day and time	2003-12
7	2003	Unknown month, day, and time	2003

为什么选择基于SDTM?

- 从EDC数据到SDTM数据标准化过程中，往往会发现一些数据问题，或者是逻辑不太对的地方，比如SDTM.AE.EPOCH，会遇到AESTDTC/AEENDTC记录时间比今天Today()的时间还要晚，等等问题



那么，如何去实现呢？

基于SDTM数据集开发提高临床数据质量工具

基于SDTM数据集开发提高临床数据质量工具

- 基于逻辑列表进行宏程序的编写

	A	B	C	D	E
			SDTM Issue Checklist		
	Domain	Item No.	Issue Description	Severity	Develop language
	AE	ae_06	不良事件对研究药物采取的措施(AEACN)为剂量增加或剂量减少，吃药记录(EX.EXDOSE)没有发生变化。	High	SAS
	AE	ae_07	不良事件对研究药物采取的措施(AEACN)为剂量未改变，吃药记录(EX.EXDOSE)剂量发生变化且原因空缺(EX.EXADJ)。	High	SAS
	AE	ae_08	不良事件对研究药物采取的措施(AEACN)为永久退出治疗，不良事件的发生日期(AESTDTC)对应的下一访视(EX.VISITNUM)仍有治疗(EX.EXDOSE)。	High	SAS
	CM	cm_01	伴随用药的结束日期(CMENDTC)晚于当前日期(TODAY())	High	SAS
	CM	cm_02	伴随用药的开始日期年份(CMSTDTC)早于出生日期年份(BRTHDTC)	High	SAS
	CM	cm_03	伴随用药的结束日期年份(CMENDTC)早于出生日期年份(BRTHDTC)	High	SAS
	MH	mh_01	既往病史的结束日期(MHENDTC)晚于当前日期(TODAY())	High	SAS
	MH	mh_02	既往病史的开始日期年份(MHSTDTC)早于出生日期年份(BRTHDTC)	High	SAS
	MH	mh_03	既往病史的结束日期年份(MHENDTC)早于出生日期年份(BRTHDTC)	High	SAS
	EC	ec_01	同一受试者相同用药(EXTRT)同一单位(ECDOSU),用药剂量(ECDOSE)不在(均值+/-3*标准差）范围内	High	SAS
	EC	ec_02	受试者药物开始日期(ECSTDTC)早于该受试者知情同意书(DM.RFICDTC)	High	SAS
	EC	ec_03	受试者药物结束日期(ECENDTC)晚于当前日期(TODAY())	High	SAS
	EX	ex_01	同一受试者相同用药(EXTRT)同一单位(EXDOSU),用药剂量(EXDOSE)不在(均值+/-3*标准差）范围内	High	SAS
	EX	ex_02	受试者药物开始日期(EXSTDTC)早于该受试者知情同意书(DM.RFICDTC)	High	SAS
	EX	ex_03	受试者药物结束日期(ECENDTC)晚于当前日期(TODAY())	High	SAS
	EG	eg_01	同一受试者的同一指标(EGTEST),测量结果(EGSTRESN)不在(均值+/-3*标准差）范围内	High	SAS
	EG	eg_02	ECG检查日期(AESTDTC)早于该受试者知情同意书(DM.RFICDTC)	High	SAS
	EG	eg_03	检查结果为异常(EGORRES),有对应的临床意义判定(QNAM = EGCLSIG)	High	SAS
	LB	lb_01	同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,存在多套标准单位(LBSTRESU)	High	SAS
	LB	lb_02	同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同一单位(LBSTRESU)的检查结果(LBSTR	High	SAS
	LB	lb_03	同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同一单位(LBSTRESU)的正常范围值下限(	High	SAS
	LB	lb_04	同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同一单位(LBSTRESU)的正常范围值上限(	High	SAS
	LB	lb_05	实验室样本采集日期(LBDTC)小于受试者知情同意书日期(DM.RFICDTC)	High	SAS
	LB	lb_06	实验室测量值不为空(LBSTRESN)且在上下限范围之外(LBSTNRLO/LBSTNRHI), 参考范围标识(LBNRIND)没有值	High	SAS
	LB	lb_07	原始值(LBORRES)与标准值(LBSTRESN)的转换因子与单位转换关系(LBORRESU/LBSTRESU)严重不符，请确认	High	SAS
	LB	lb_08	原始下限(LBORNRLO)与标准下限(LBSTNRLO)的转换因子与单位转换关系(LBORRESU/LBSTRESU)严重不符，请确认	High	SAS
	LB	lb_09	原始上限(LBORNRHI)与标准上限(LBSTNRHI)的转换因子与单位转换关系(LBORRESU/LBSTRESU)严重不符，请确认	High	SAS

Issue Checklist | Retired items | Revision History

基于SDTM数据集开发提高临床数据质量工具

SDTM Issue Checklist				
Domain	Item No.	Issue Description	Severity	Develop language
AE	ae_04	不良事件的开始日期(AESTDTC)早于该受试者知情同意书日期(DM.RFICDTC)	High	SAS
AE	ae_05	不良事件的结束日期(AEENDTC)早于该受试者知情同意书日期(DM.RFICDTC)	High	SAS
EC	ec_02	受试者药物开始日期(ECSTDTC)早于该受试者知情同意书(DM.RFICDTC)	High	SAS
EX	ex_02	受试者药物开始日期(EXSTDTC)早于该受试者知情同意书(DM.RFICDTC)	High	SAS
EG	eg_02	ECG检查日期(AESTDTC)早于该受试者知情同意书(DM.RFICDTC)	High	SAS
LB	lb_05	实验室样本采集日期(LBDTC)小于受试者知情同意书日期(DM.RFICDTC)	High	SAS
VS	vs_01	生命体征采集日期(VSDTC)小于受试者知情同意书日期(DM.RFICDTC)	High	SAS

- 对于核查时间程序，可以通过SDTM的时间变量命名规则，通过宏循环批量操作

```
%let sdtm_dataset = CM / MH / EC / EX /;
%do i = 1 %to %eval(%sysfunc(count(&sdtm_dataset.,)) + 1);

    %let sd = %cmpres(%scan(&sdtm_dataset.,&i.,/));
    proc sql;
        create table c2_&sd._dm_1 as
            select distinct a.*
                          ,b.rfpendtc,b.rficdtc,b.brthdtc
                from sdtm.&sd. as a left join sdtm.dm as b
                    on a.usubjid = b.usubjid
        ;
    quit;

    data c2_&sd._dm_r;
        set c2_&sd._dm_1;
        if &sd.STDTC < RFICDTC then do;
            Variables = "USUBJID/&SD.STDTC/RFICDTC";;
            Value = catx("/",USUBJID,&SD.STDTC,RFICDTC);
            Seq = &SD.Seq;
            Domain = "&SD.";
            ID = "&SD._0x";
            Message = "开始日期(&SD.STDTC)早于该受试者知情同意书日期(RFICDTC)。";
            Severity = "High";
            output;
        end;
    run;
%end;
```

基于SDTM数据集开发提高临床数据质量工具

同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同一单位(LBSTRESU)的检查结果(LBSTRESN)不在(均值+/-3*标准差）范围内
同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同一单位(LBSTRESU)的正常范围值下限(LBSTNRLO)不在(均值+/-3*标准差）范围内
同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同一单位(LBSTRESU)的正常范围值上限(LBSTNRHI)不在(均值+/-3*标准差）范围内

```
data c2_lb_1;
    set sdtm.lb;
    array a(*) $200 LBCAT LBTEST LBMETHOD LBSPEC LBSTRESU;
    _Sort_ = _N_;
run;

proc sort data=c2_lb_1;
    by LBCAT LBTEST LBMETHOD LBSPEC LBSTRESU;
run;

ods trace on ;
ods exclude all;
ods output Summary = Summary;
proc means data=c2_lb_1 mean std ;
    by LBCAT LBTEST LBMETHOD LBSPEC LBSTRESU;
    var lbstresn;
run;
ods exclude none ;
ods trace off;
ods output close;

proc sql;
    create table c2_lb_2 as
        select a.*
             ,b.LBSTRESN_Mean,b.LBSTRESN_StdDev
            from c2_lb_1 as a
                left join Summary as b
                    on a.LBCAT = b.LBCAT and a.LBTEST = b.LBTEST and a.LBMETHOD = b.LBMETHOD and a.LBSPEC = b.LBSPEC and a.LBSTRESU = b.LBSTRESU
                order by _Sort_
    ;
quit;

data Issue03_LB ;
    set c2_lb_2;
    if ^missing(LBSTRESN) then do;
        if LBSTRESN > LBSTRESN_Mean + 3*LBSTRESN_StdDev or LBSTRESN < LBSTRESN_Mean - 3*LBSTRESN_StdDev then do;
            output;
        end;
    end;
run;
```

- 为了确保实验室的数据录入正确，避免手误录入错误，可以使用proc mean，计算出极端值出来，交给DM同事发送质疑进行核查

基于SDTM数据集开发提高临床数据质量工具

- 为了减少宏参数的填写，程序上面可以写得更加智能，比如添加自动抓取逻辑路径功能，这样核查报告就能自动生成到数据集所存放的物理逻辑库下面。

```
%** Get Path;

data _Null_;
    Path = pathname("SDTM");
    Dattim = strip(put(datetime(), E8601DT.));;
    call symput("OutputPath",strip(Path));
    call symput("Dattim",strip(tranwrd(Dattim,":","-")));
run;

goptions device="ACTXIMG";
ods _all_ close;
ods excel(id=one) file="&OutputPath.\SDTM Validtion-&Dattim..xlsx" options(flow="tables, text");

    %** Sheet1: Coder Operating Sheet;
    ods excel(id=one) options(sheet_name="Issue Summary" autofilter="all"
            frozen_headers="1");

    proc report nowd data=issue split="$"
        style(header)={tagattr="wrap:no" just=center background=lightblue color=black}
        style(column)={tagattr="format:@"};
        column(Domain Seq Variables Value ID Message Severity); %** affect the display order;
    run;
```

基于SDTM数据集开发提高临床数据质量工具

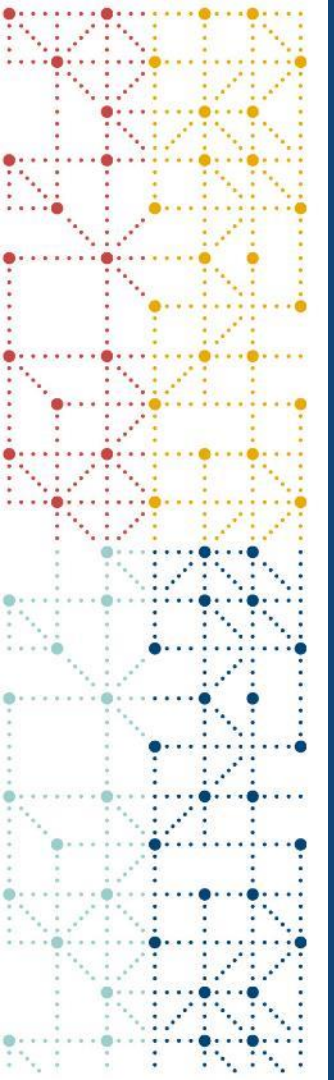
```
%**set path;
%let _pgpath_ = \\192.168.123.8\z\studies\macro-develop\0-batchrun-rb\10x-dmpr\program\      ;
%sysexec cd &_pgpath_.;
%sysexec setenv PWD &_pgpath_.;

%inc "..\init.sas";

%check_sdtm;
```

最后把工具打包成一个宏，可以贴合项目环境进行运行，生成核查报告。

Domai	Se	Variables	Value	ID	Message
AE	34	USUBJID/AEDECOD/AESTDTC/AEENDTC	-01004/Platelet count decreased/2020-12-15/2021-01-14	ae_01	同一个受试者(USUBJID)，相同的不良事件(AEDECOD)，开始日期和结束日期有交叉。
AE	35	USUBJID/AEDECOD/AESTDTC/AEENDTC	-01004/Platelet count decreased/2020-12-15/2021-01-14	ae_01	同一个受试者(USUBJID)，相同的不良事件(AEDECOD)，开始日期和结束日期有交叉。
AE	76	USUBJID/AEDECOD/AESTDTC/AEENDTC	-01006/Platelet count decreased/2022-02-04/2022-02-08	ae_01	同一个受试者(USUBJID)，相同的不良事件(AEDECOD)，开始日期和结束日期有交叉。
AE	77	USUBJID/AEDECOD/AESTDTC/AEENDTC	-01006/Platelet count decreased/2022-02-04/2022-02-21	ae_01	同一个受试者(USUBJID)，相同的不良事件(AEDECOD)，开始日期和结束日期有交叉。
AE	97	USUBJID/AEDECOD/AESTDTC/AEENDTC	-01006/White blood cell count decreased/2021-09-02/2021-09-16	ae_01	同一个受试者(USUBJID)，相同的不良事件(AEDECOD)，开始日期和结束日期有交叉。
AE	109	USUBJID/AEDECOD/AESTDTC/AEENDTC	-01006/White blood cell count decreased/2021-09-02/2021-09-16	ae_01	同一个受试者(USUBJID)，相同的不良事件(AEDECOD)，开始日期和结束日期有交叉。
AE	114	USUBJID/AEDECOD/AESTDTC/AEENDTC	-01006/White blood cell count decreased/2022-02-04/2022-02-08	ae_01	同一个受试者(USUBJID)，相同的不良事件(AEDECOD)，开始日期和结束日期有交叉。
AE	115	USUBJID/AEDECOD/AESTDTC/AEENDTC	-01006/White blood cell count decreased/2022-02-04/2022-03-18	ae_01	同一个受试者(USUBJID)，相同的不良事件(AEDECOD)，开始日期和结束日期有交叉。
AE	116	USUBJID/AEDECOD/AESTDTC/AEENDTC	-01006/White blood cell count decreased/2022-02-08/2022-02-17	ae_01	同一个受试者(USUBJID)，相同的不良事件(AEDECOD)，开始日期和结束日期有交叉。
AE	117	USUBJID/AEDECOD/AESTDTC/AEENDTC	-01006/White blood cell count decreased/2022-02-25/2022-03-02	ae_01	同一个受试者(USUBJID)，相同的不良事件(AEDECOD)，开始日期和结束日期有交叉。



实际项目运行实例

实际项目运行实例

在SDTM Dryrun阶段，项目中会把多套单位转化成申办方要求的标准单位。

	F	G	H	I	J	K	L
	LBTEST	LBTESTC	LBSPEC	Original_Unit	LBORRESU	LBSTRESU	Conversion Factor
	Thyroxine, Free	T4FR		pmol/L	pmol/L	ng/L	0.777
	Thyroxine, Free	T4FR		ng/dl	ng/dL	ng/L	10
	Thyroxine, Free	T4FR		ng/dL	ng/dL	ng/L	10
	Thyroxine, Free	T4FR		ng/mL	ng/mL	ng/L	1000
	Thyroxine, Free	T4FR		pg/mL	pg/mL	ng/L	1
	Thyroxine, Free	T4FR					

实际项目运行实例

SDTM Check

在SDTM Check阶段，运行自动核查工具，检查出多个离群值，此时可以把问题反馈给DM同事进行质疑核查。

同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同一单位(LBSTRESU)的检查结果(LBSTRESN)不在(均值+/-3*标准差）范围内
同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同一单位(LBSTRESU)的正常范围值下限(LBSTNRLO)不在(均值+/-3*标准差）范围内
同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同一单位(LBSTRESU)的正常范围值上限(LBSTNRHI)不在(均值+/-3*标准差）范围内

A	B	C	D	E	F
域名缩写	Se	Variables	Value	ID	Message
LB	94	USUBJID/LBTEST/LBSPEC/LBSTRESN/LBSTRESU	33140300027-S10001/游离甲状腺素/1080/ng/L	LB_01	同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同结果(LBSTRESN)不在(均值+/-3*标准差）范围内
LB	95	USUBJID/LBTEST/LBSPEC/LBSTRESN/LBSTRESU	33140300027-S10001/游离甲状腺素/1310/ng/L	LB_01	同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同结果(LBSTRESN)不在(均值+/-3*标准差）范围内
LB	137	USUBJID/LBTEST/LBSPEC/LBSTRESN/LBSTRESU	33140300027-S10002/游离甲状腺素/950/ng/L	LB_01	同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同结果(LBSTRESN)不在(均值+/-3*标准差）范围内
LB	138	USUBJID/LBTEST/LBSPEC/LBSTRESN/LBSTRESU	33140300027-S10002/游离甲状腺素/920/ng/L	LB_01	同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同结果(LBSTRESN)不在(均值+/-3*标准差）范围内
LB	139	USUBJID/LBTEST/LBSPEC/LBSTRESN/LBSTRESU	33140300027-S10002/游离甲状腺素/620/ng/L	LB_01	同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同结果(LBSTRESN)不在(均值+/-3*标准差）范围内
LB	137	USUBJID/LBTEST/LBSPEC/LBSTRESN/LBSTRESU	33140300027-S13001/游离甲状腺素/980/ng/L	LB_01	同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同结果(LBSTRESN)不在(均值+/-3*标准差）范围内
LB	138	USUBJID/LBTEST/LBSPEC/LBSTRESN/LBSTRESU	33140300027-S13001/游离甲状腺素/1030/ng/L	LB_01	同一分类(LBCAT)同一检查项(LBTEST)同一方法(LBMETHOD)同一样本类型(LBSPEC)下,同结果(LBSTRESN)不在(均值+/-3*标准差）范围内

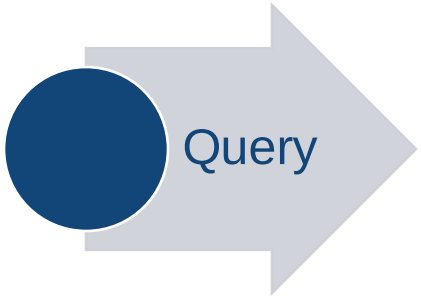
实际项目运行实例

TFL Dryrun

在SDTM Check阶段过后，项目组成员在做TFL Dryrun时，会对该指标的结果进行原因说明，并把问题反馈给统计师。

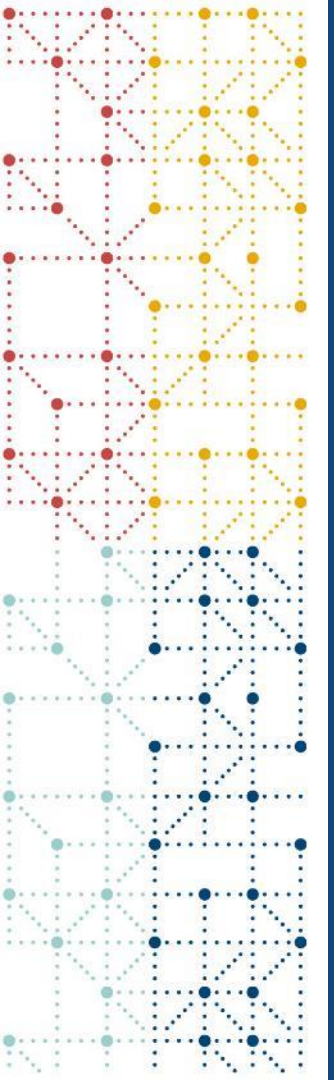
Thyroxine, free (ng/L)			
Baseline			
N	186		184
Mean (SD)	17.521 (68.8002)		23.231 (105.9862)
Median (Q1, Q3)	12.370 (10.412, 14.071)		12.281 (10.200, 13.823)
Min, Max	6.45, 950.00		7.11, 1080.00

实际项目运行实例



在数据管理核查中，这种错误很难被发现发现，但是在SDTM阶段中，加上自动化工具核查功能，进一步提高了数据的准确性。

当前值	正确值	备
[游离甲状腺素] 筛选期 1.08 ng/mL 第6周/ET 1.31 ng/mL	[游离甲状腺素] 筛选期 1.08 ng/dl 第6周/ET 1.31 ng/dl	[2 错 [2 锁 [2 共
[游离甲状腺素] 筛选期 0.95 ng/mL 第3周 0.92 ng/mL 第6周/ET 0.62 ng/mL	[游离甲状腺素] 筛选期 0.95 ng/dl 第3周 0.92 ng/dl 第6周/ET 0.62 ng/dl	[2 误 [2 锁 [2 程
[总甲状腺素] 筛选期 8.84 ug/mL 第3周 9.54 ug/mL [游离甲状腺素] 筛选期 0.98 ng/mL 第3周 1.03 ng/mL	[总甲状腺素] 筛选期 8.84 ug/dL 第3周 9.54 ug/dL [游离甲状腺素] 筛选期 0.98 ng/dL 第3周 1.03 ng/dL	[2 误 总 [2 锁 [2 程



总结

总结

基于SDTM数据集开发提高临床数据质量工具：

- 能够提高数据核查的质量；
- SDTM标准，更加友好地让程序员实现自动化操作；

Thank You!