



Data Anonymization

Providing clinical trial data to outside researchers

PhUSE SDE, Mumbai

14March2015

Santhosh Lyathakula



Outline

Topics covered in today presentation

The Regulatory Background

The Objectives – Protection and Analyses

Protection – Deidentification or Anonymization?

Novartis approach

Overview of the anonymization process

Modes of anonymization

Results: Real data Vs. Anonymized data

QA

Patient Confidentiality

Masking or removing data is not as easy as one could think



Masking or removing information needs to be done in a consistent way and is all about finding the right balance between protecting patient privacy while maintaining data integrity and reusability

The Regulatory Background

- EMA requirement to provide clinical trial data to outside researchers to perform independent analyses
 - Supports EMA objective to make decision making transparent (Policy 0043)
 - Policy 0070 Publication and Access to Clinical Trial Data
- There are also objectives in the USA to be more transparent, expressed in the PhRMA Principles
- Other EU and USA regulations require the protection of personal data (e.g. Regulation (EC) No 45/2001 and Directive 95/46/EC and HIPPA) and to stay within the scope of informed consent
- **EMA Position**: Clinical trial data is not commercial, confidential information.

The Objectives – Protection and Analyses

- OBJECTIVE 1: Data privacy protection
 - Protect all the data values that can identify a patient
- OBJECTIVE 2: To create a database that can be used for meaningful analyses and for the testing of internal systems.
 - Keep the data values that do not identify a patient – test results, etc.
 - Support the ability to merge data
 - Know that the same person with these lab tests had these AEs
 - Perform chronological analyses such as time to onset, duration, concurrency, etc.

Protection – Deidentification or Anonymization?

- Two terms have come into the discussion
 - Deidentification – protect data privacy and allow the capability to re-identify the data
 - Anonymization – protect data privacy and prohibit the capability to re-identify the data
- Novartis has decided to anonymize data
- The regulations may be subject to interpretation, such that deidentification might be allowed.
 - Data privacy regulations apply to all industries, not just pharma. Some of these industries are *required to delete* identifying data as soon as they are finished with it. These non-pharma industries can retain the data only if they anonymize it. Pharma is *required to keep* the real data, so deidentification may be allowed.

Novartis approach

The anonymization process

A process has been developed to “anonymize” clinical study datasets through use of a standard macro and a set of data definition and ***anonymization mode*** attributes.



These attributes are passed to the anonymization macro via a ***SAS definition dataset*** and may come from two sources.



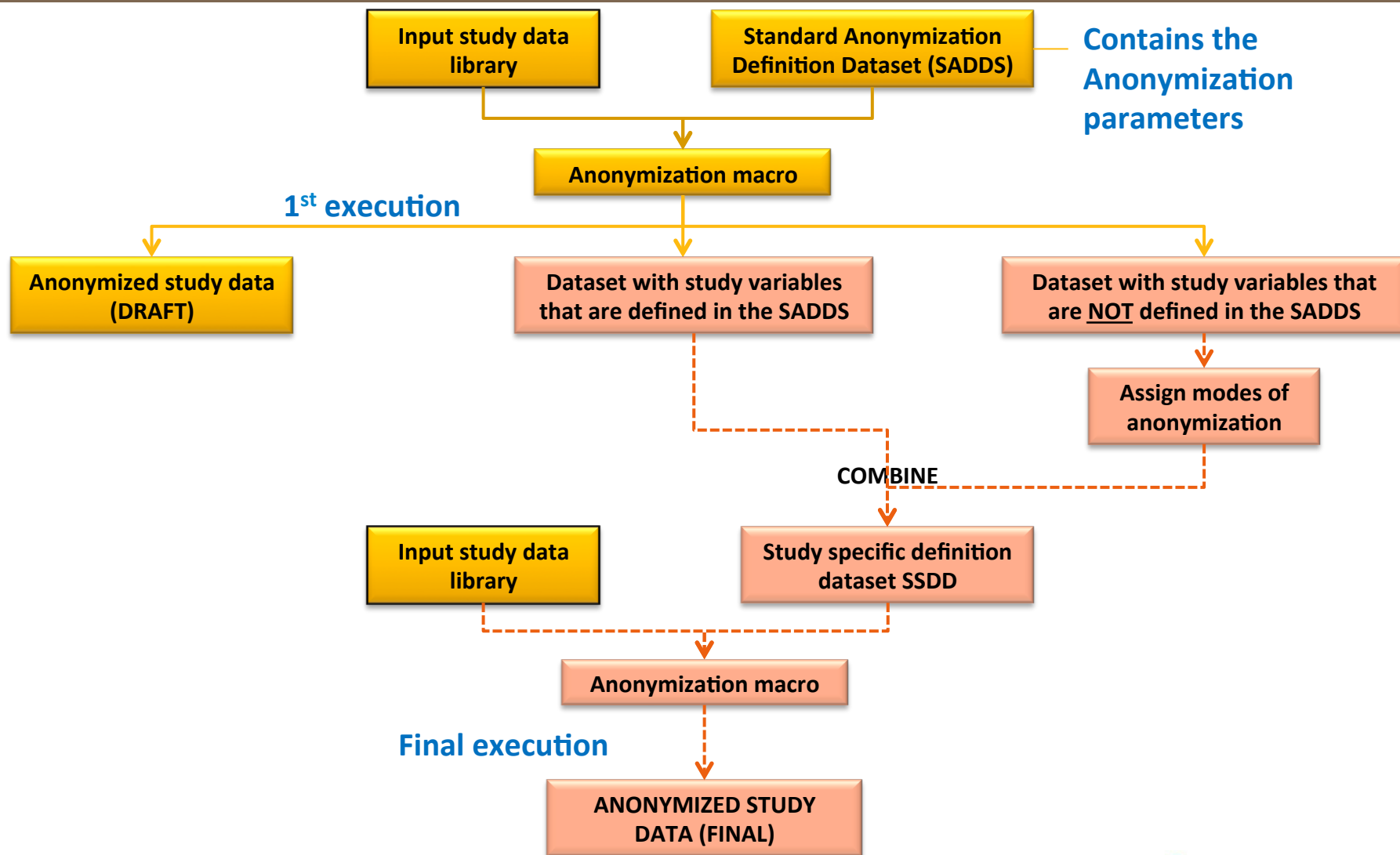
The **Standard** Anonymization Definition Dataset (SADDs) is used across studies and is a repository of standard data definition and mode attributes.



The **Study Specific** Definition Dataset is used when there are variables collected in the **study** datasets that haven't been defined in the SADDs.

Overview of Anonymization process

Proposed solution for creating a Study specific definition dataset



A High-level view of the Definition Dataset

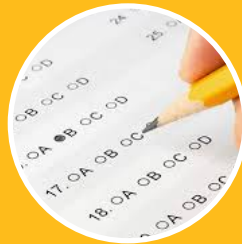
DATASET	VARIABLE	LABEL	MODE	identifier_	date_type	full_date	TYPE	FORMAT
AEFF2EVT	ACSDT	Date of 1st Hosp. for ACS	DATE	date	sasdate		1	DATE
ACMDATC	CMDEND10	Concomitant med. end date (Oracle date)	DATE	datetime	sasdatetime		1	DATETIME
AAEV	AEVEND1D	Adverse event end (date)	DROP				2	\$
AAEV	AEVSTT1D	Adverse event start (date)	DROP				2	\$
AAEV	AGECL65	Age group (<65,>=65)	DROP				1	AGE1F_
ACMD	AGECL65	Age group (<65,>=65)	DROP				1	AGE1F_
ACMDATC	AGECL65	Age group (<65,>=65)	DROP				1	AGE1F_
ACMP	ACTTRTC	Actual treatment code	NONE				2	\$
ACOM	ACTTRTC	Actual treatment code	NONE				2	\$
ADAR	ACTTRTC	Actual treatment code	NONE				2	\$
AADJ	CTR1N	Center number	TRANSLATE				1	
AAEV	CTR1N	Center Number	TRANSLATE				1	
ABIO	SID1A	Subject Identifier	TRANSLATE				2	\$
ABKG	SID1A	Subject Identifier	TRANSLATE				2	\$
ACMPDTH	SID1A	Subject Identifier	TRANSLATE				2	\$

- ✓ List of all variables from all datasets to be anonymized
- ✓ This is the place where the modes of anonymization are entered
- ✓ The Definition Dataset is standardized at the Division level and is maintained through a change management system (version history and approval process)

Process flow



Novartis
Anonymization
Guidelines



Anonymization
Mode
Selection



Programming
Considerations



Output
Anonymized
data



Novartis Global Data Anonymization Standards

Contents	
1. Introduction	1
2. General Approach	1
3. Removing Personally Identifiable Information (PII)	2
3.1. PII	2
3.2. Identifiers	3
3.3. Free Text Verbatim Terms	3
3.4. Date of Birth	3
3.5. Other Dates	3
3.6. Other PII	4
4. Remnants	4
5. Example	5
6. Reference List	5

1 Introduction

Patient-level data collected in Novartis clinical trials will be anonymized according to the standards set forth in this document. These standards will ensure compliance with current privacy laws and regulatory guidance while allowing data to be shared with researchers. There are a number of data elements enumerated in the "Privacy Rule" under the Health Insurance Portability and Accountability Act (HIPAA) of 1996 and other guidance from European General Data Protection Regulation which can be used to identify individuals. The process of anonymizing can be thought of as permanently removing the ability to use any of these elements to identify individual participants. Direct and indirect identifiers are removed thereby making it unlikely to allow any individual to be identified by combining data. Adherence to the framework of these standards will minimize the risks of encroaching on the privacy and confidentiality of research participants.

2 General Approach



Novartis
Anonymization Guidelines



Modes of anonymization

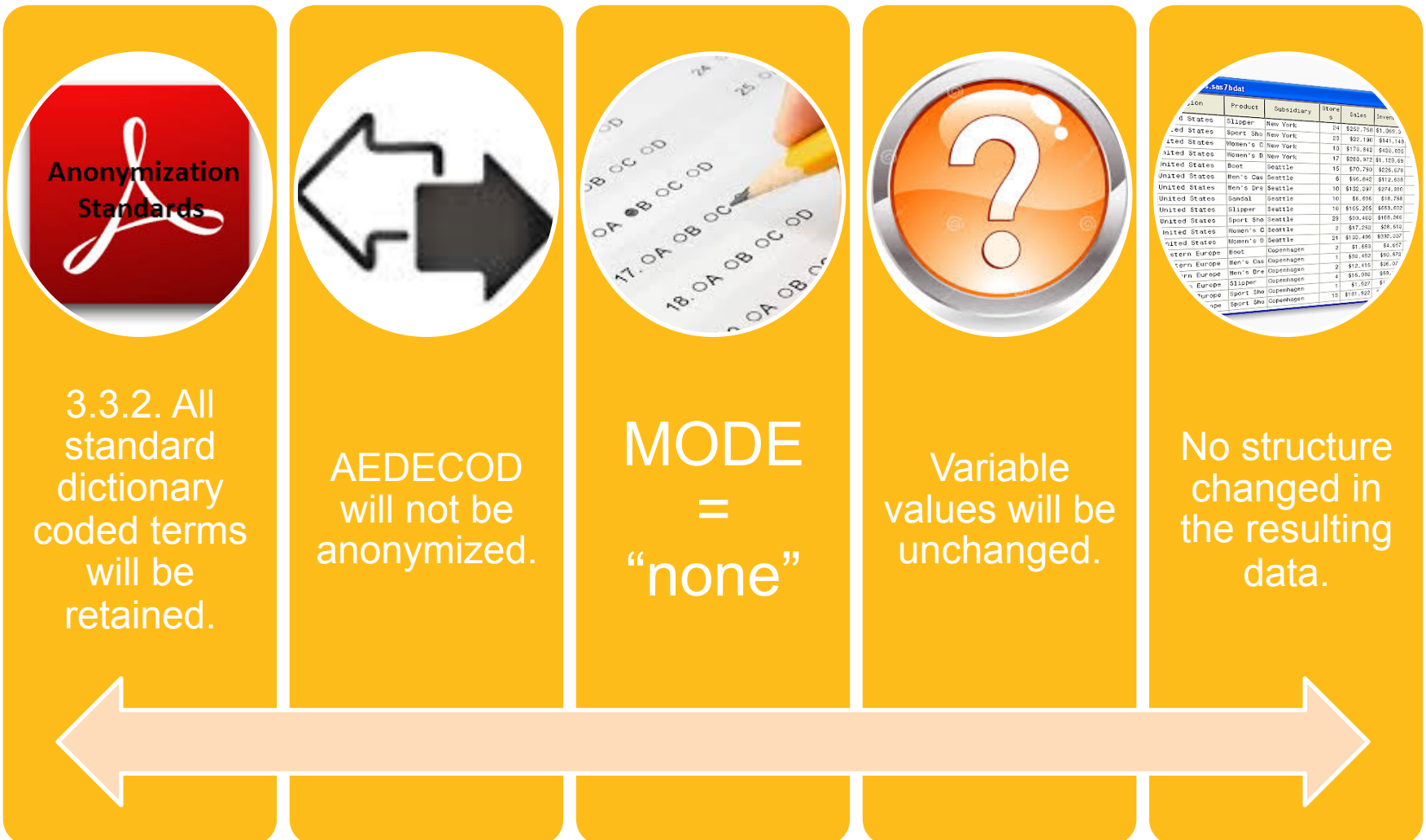
Walkthrough and examples

Following are the modes of Anonymization:

- none
- missing
- drop
- ageint
- date
- translate

MODE = “none”

Straight copy of the variable



Region	Product	Subsidiary	Store	Sales	Inventory
United States	Slipper	New York	24	\$222,750	\$1,200,500
United States	Sport Sho	New York	23	\$22,794	\$141,145
United States	Women's C	New York	13	\$173,842	\$420,402
United States	Women's B	New York	17	\$283,372	\$1,120,426
United States	Boat	Seattle	15	\$70,790	\$226,679
United States	Men's Cas	Seattle	6	\$95,642	\$912,638
United States	Men's Dre	Seattle	10	\$125,297	\$274,580
United States	Sandal	Seattle	10	\$6,698	\$18,788
United States	Slipper	Seattle	10	\$164,266	\$653,402
United States	Sport Sho	Seattle	29	\$50,440	\$69,384
United States	Women's C	Seattle	7	\$17,250	\$28,432
United States	Women's B	Seattle	21	\$125,496	\$370,227
Western Europe	Boat	Copenhagen	2	\$1,663	\$4,427
Western Europe	Men's Cas	Copenhagen	1	\$36,402	\$60,579
Western Europe	Men's Dre	Copenhagen	2	\$72,415	\$26,37
Western Europe	Slipper	Copenhagen	4	\$16,396	\$55
Western Europe	Sport Sho	Copenhagen	1	\$1,627	\$1
Western Europe	Sport Sho	Copenhagen	10	\$101,927	\$1

MODE SELECTION = "none"

Original Data

SUBJID	AETERM	AESTDTC	AEDECOD	AEENDTN	INV	ORIGSITE
1001001	Broken Crown on...	2013-04-1...	DEVICE BREAK...	2013-04-10T00:0...	308667C@CA	1011
1001002	pharyngitis [due t...	2012-11-...	PHARYNGITIS	2012-12-02T00:0...	308679C@CA	1017
1001003	pharyngitis [due t...	2012-11-...	PHARYNGITIS	2012-12-02T00:0...	308679C@CA	1017
1001007	stomach virus	2012-12-...	GASTROENTER	2012-12-18T00:0...	127996C@CA	1023
1001009	vomiting	2012-09-29	VOMITING	2012-10-03T00:0...	277526C@CA	1025

Definition Data set

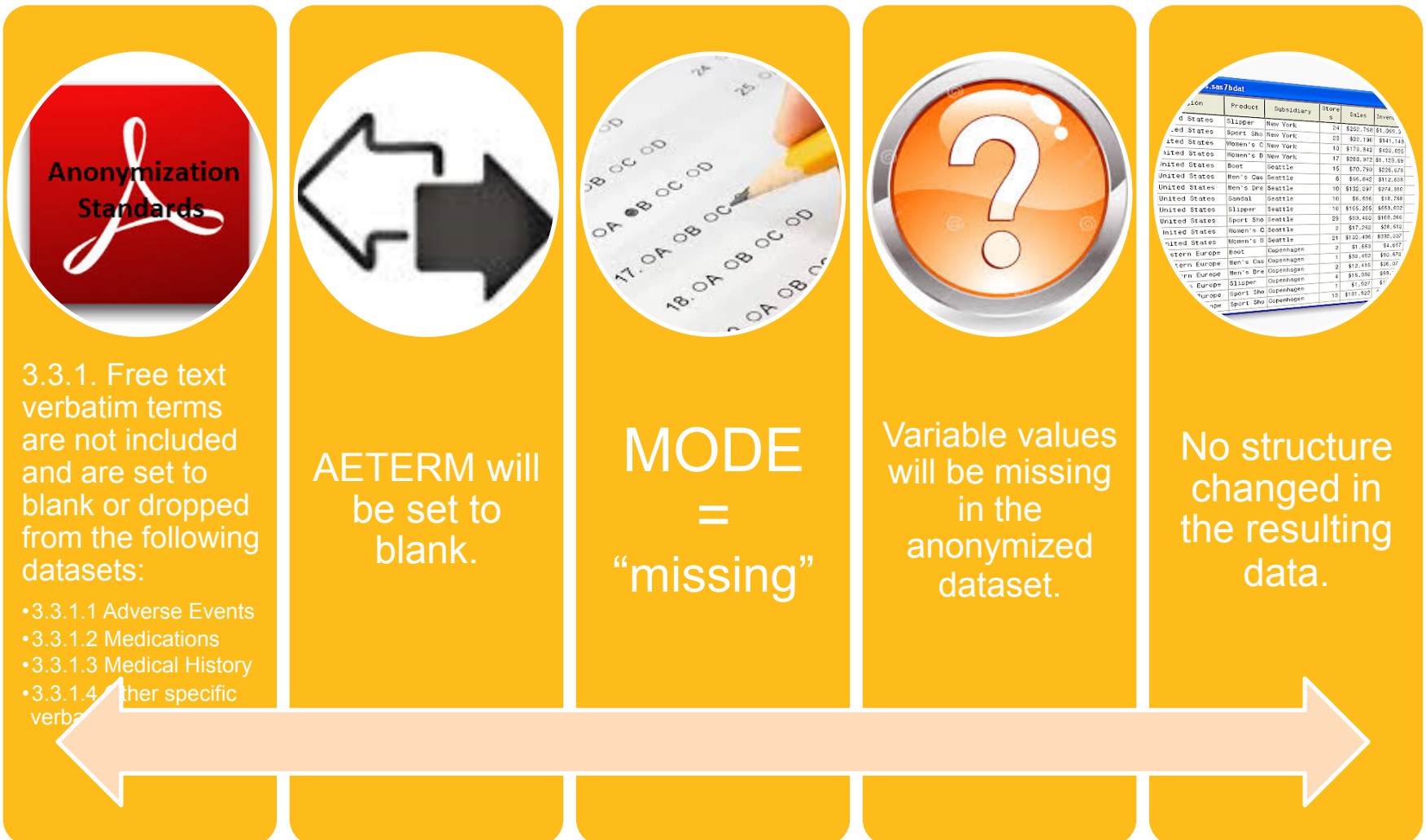
table	column	mode	identifier_type	date_type	full_date	LABEL
AE	AEDECOD	none				AE Term : MedD...
AE	AEEDT	date	date	sasdate		Imputed AE end ...
AE	AEEDT	date	datec	is8601da.		
AE	AEENDTC	date	datetime	sasdatetime		End Date/Time o...
AE	AEENDTN	date	datec	is8601da.		Start Date/Time ...
AE	AEENDTN	date				Reported Term fo...

Anonymized Data

SUBJID	AETERM	AESTDTC	AEDECOD	AEENDTN	INV	ORIGSITE
nnnnnn1		2323-07-1...	DEVICE BREAK...	2013-04-10T00:0...	nnnnnnX@11	nnn5
nnnnnn2		2323-03-...	PHARYNGITIS	2012-12-02T00:0...	nnnnnnX@12	nnn6
nnnnnn3		2323-03-...	PHARYNGITIS	2012-12-02T00:0...	nnnnnnX@12	nnn6
nnnnnn4		2323-03-2...	GASTROENTER	2012-12-18T00:0...	nnnnnnX@X4	nnnB
nnnnnn5		2323-01-06	VOMITING	2012-10-03T00:0...	nnnnnnX@XD	nnnC

MODE = “missing”

Keeping variable but removing its contents



MODE SELECTION = "missing"

Original Data

SUBJID	AETERM	AESTDTC	AEDECOD	AEENDTN	INV	ORIGSITE
100100	Broken Crown on...	2013-04-10	DEVICE BREAK...	2013-04-10T00:0...	308667C@CA	1011
100101	pharyngitis [due t...	2012-11-25	PHARYNGITIS	2012-12-02T00:0...	308679C@CA	1017
100102	pharyngitis [due t...	2012-11-25	PHARYNGITIS	2012-12-02T00:0...	308679C@CA	1017
100103	stomach virus	2012-12-17	GASTROENTER...	2012-12-18T00:0...	127996C@CA	1023
100109	vomiting	2012-09-29	VOMITING	2012-10-03T00:0...	277526C@CA	1025

Definition Data set

table	column	mode	identifier_type	date_type	full_date	LABEL
AE	AEDECOD	none				AE Term : MedD...
AE	AEEDT	date	date	sasdate		Imputed AE end ...
AE	AEENDTC	date	datec	is8601da.		
AE	AEENDTN	date	datetime	sasdatetime		End Date/Time o...
AE	AESTDTC	date	datetime	sasdatetime		Start Date/Time ...
AE	AETERM	missing	datec	is8601da.		Reported Term fo...

Anonymized Data

SUBJID	AETERM	AESTC	AEDECOD	AEENDTN	INV	ORIGSITE
nnnnnn1		2323-8	DEVICE BREAK...	2323-07-18T00:0...	nnnnnnX@11	nnn5
nnnnnn2		2323-0	PHARYNGITIS	2323-03-11T00:0...	nnnnnnX@12	nnn6
nnnnnn3		2323-0	PHARYNGITIS	2323-03-11T00:0...	nnnnnnX@12	nnn6
nnnnnn4		2323-0	GASTROENTER...	2323-03-27T00:0...	nnnnnnX@X4	nnnB
nnnnnn5		2323-06	VOMITING	2323-01-10T00:0...	nnnnnnX@XD	nnnC

MODE = “drop”

Dropping variable

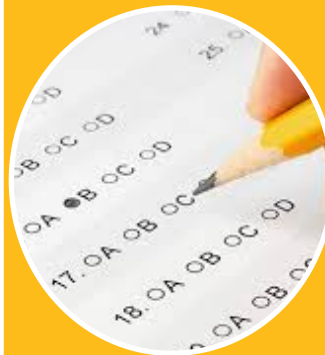


3.4 Date of Birth

- Information relating to a research participant's date of birth and identification of specific ages above 89 may compromise anonymity.
- Date of birth is dropped and ages above 89 are aggregated into a single category of “90 or older”.



BRTHDTC
will be
dropped in
the
anonymized
dataset.



MODE
=
“drop”



Variable will
be dropped
in the
anonymized
dataset.

A screenshot of a data table with columns: Location, Product, Subsidiary, Store, Sales, and Inventory. The table lists various products like Slippers, Sport Shoes, and Boots across different locations and subsidiaries.

Structure will
be changed
by dropping
a column in
the resulting
data.

MODE SELECTION = "drop"

Original
Data

ORIGSITE	AEENDTC	AGE	AEEDT	BRTHTDC
1011	2013-04-10T00:0...	47	2013-07-10	1965-04-08T18:00
1017	2012-12-02T00:0...	45	2012-02-02	1967-11-06T18:00
1017	2012-12-02T00:0...	97	2012-02-02	1967-11-06T18:00
1023	2012-12-18T00:0...	31	2012-10-18	1981-06-04T18:00
1025	2012-10-03T00:0...	49	2012-10-18	1963-02-26T18:00

Definition
Data set

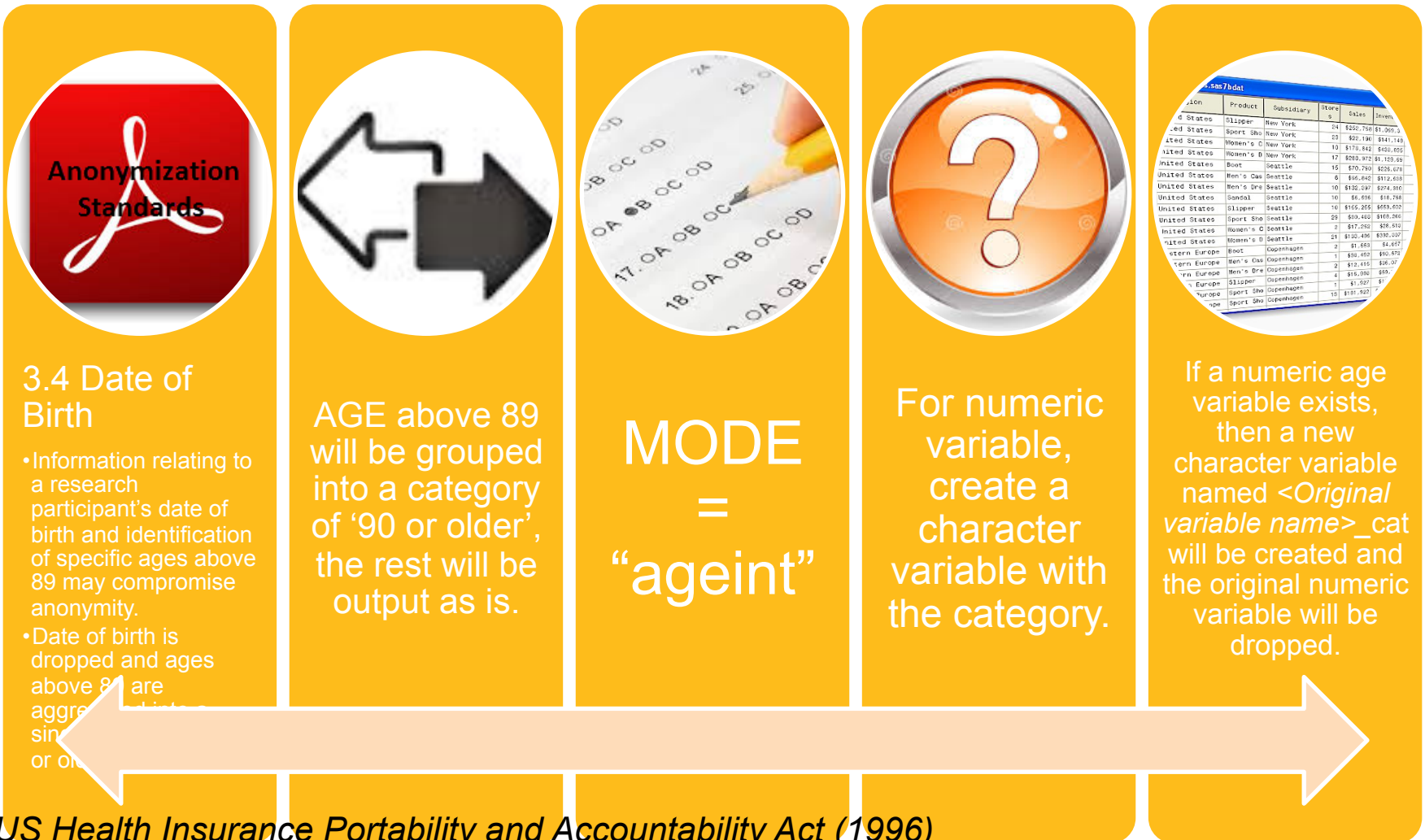
table	column	mode	identifier_type	date_type	full_date	LABEL
AE	AEDECOD	none				AE Term : MedDRA Preferred Term
AE	AEEDT	date	date	sasdate		Imputed AE end date
AE	AEENDTC	date	datec	is8601da.		
AE	AEENDTN	date	datetime	sasdatetime		End Date/Time of Adverse Event
AE	AESTDTC	date	datec	is8601da.		Start Date/Time of Adverse Event
AE	AETERM	missing				Reported Term for the Adverse Event
AE	AGE	ageint				Age in AGEU at RFSTDTC
AE	BRTHTDC	drop				

Anonymized
Data

ORIGSIT	AEENDTC	AEEDT	age_cat
nnn5	2328-04-24	2328-04-24	47
nnn6	2327-12-17	2327-12-17	45
nnn6	2327-12-17	2327-12-17	90 or older
nnnB	2328-01-02	2328-01-02	31
nnnC	2327-10-18	2327-10-18	49

MODE = “ageint”

Grouping ages above 89years (HIPAA* requirement)



*US Health Insurance Portability and Accountability Act (1996)

MODE SELECTION = "ageint"

Original
Data

INV	ORIGSITE	AEENDTC	AGE	AEEDT
308667C@CA	1011	2013-04-10T...	47	2013-04-10
308679C@CA	1017	2012-12-02T0...	45	2012-12-02
308679C@CA	1017	2012-12-02T0...	97	2012-12-02
127996C@CA	1023	2012-12-18T0...	31	2012-12-18
277526C@CA	1025	2012-10-03T0...	49	2012-10-03

Definition
Data set

table	column	mode	identifier_type	date_type	full_date	LABEL
AE	AEDECOD	none				AE Term : MedD...
AE	AEEDT	date	date	sasdate		Imputed AE end ...
AE	AEENDTC	date	datec	is8601da.		
AE	AEENDTN	date	datetime	sasdatetime		End Date/Time o...
AE	AESTDTC	date	datec	is8601da.		Start Date/Time ...
AE	AETERM	missing				Reported Term fo...
AE	AGE	ageint				Age in AGEU at ...

Anonymized
Data

AEENDTC	AEEDT	BRTHTC	age_cat
2323-07-18	2323-07-18	2275-07-15T18:00:00	47
2323-03-11	2323-03-11	2278-02-11T18:00:00	45
2323-03-11	2323-03-11	2278-02-11T18:00:00	90 or older
2323-03-27	2323-03-27	2291-09-10T18:00:00	31
2323-01-10	2323-01-10	2273-06-03T18:00:00	49

MODE = “date”

Offset dates



MODE SELECTION = "date"

Original
Data

INV	ORIGSITE	AEENDTC	AGE	AEEDT
308667C@CA	1011	2013-04-10T00:0...	47	2013-04-10
308679C@CA	1017	2012-12-02T00:0...	45	2012-12-02
308679C@CA	1017	2012-12-02T00:0...	97	2012-12-02
127996C@CA	1023	2012-12-18T00:0...	31	2012-12-18
277526C@CA	1025	2012-10-03T00:0...	49	2012-10-03

Definition
Data set

table	column	mode	identifier_type	date_type	full_date	LABEL
AE	AEDECOD	none				AE Term : MedD...
AE	AEEDT	date	date	sasdate		Imputed AE end ...
AE	AEENDTC	date	datec	is8601da.		
AE	AEENDTN	date	datetime	sasdatetime		End Date/Time o...
AE	AESTDTC	date	datec	is8601da.		Start Date/Time ...
AE	AETERM	missing				Reported Term fo...
AE		age				Age in AGEU at ...

Anonymized
Data

AEENDTC	AEEDT	BRTHTDC	age_cat
2323-07-18	2323-07-18	2275-07-15T18:00:00	47
2323-03-11	2323-03-11	2278-02-11T18:00:00	45
2323-03-11	2323-03-11	2278-02-11T18:00:00	90 or older
2323-03-27	2323-03-27	2291-09-10T18:00:00	31
2323-01-10	2323-01-10	2273-06-03T18:00:00	49

MODE = “translate”

Masking data

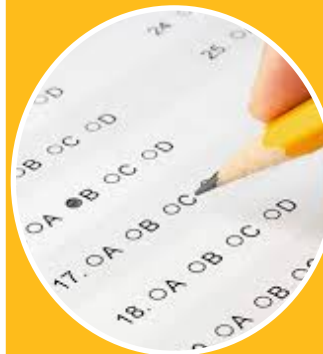


3.2 Identifiers Change the real value to a de-identified value in a consistent manner so that the value in one instance of the variable is consistent with the value in the same variable across other datasets.

- 3.2.1. The investigator number is re-coded or set to blank for each investigator. The investigator name is set to blank.
- 3.2.2. Each participant is given a new subject identifier.



SUBJID will be translated



MODE = “translate”



For consistent translation of values between datasets of a study or between datasets of multiple studies, all study data must be anonymized at the same time to use the same translation table.

Location	Product	Subsidiary	Store	Sales	Inventory
United States	Slippers	New York	24	\$222,750	\$1,000,000
United States	Sport Shoes	New York	23	\$22,794	\$141,148
United States	Women's Bras	New York	13	\$172,842	\$420,402
United States	Women's Bras	New York	17	\$283,372	\$1,120,426
United States	Boots	Seattle	15	\$70,790	\$226,679
United States	Men's Bras	Seattle	6	\$95,642	\$912,638
United States	Men's Bras	Seattle	10	\$125,297	\$274,940
United States	Sandals	Seattle	10	\$6,698	\$18,748
United States	Slippers	Seattle	10	\$166,266	\$683,432
United States	Sport Shoes	Seattle	29	\$50,440	\$68,384
United States	Women's Bras	Seattle	2	\$17,250	\$28,432
United States	Women's Bras	Seattle	21	\$125,496	\$370,227
United States	Boots	Copenhagen	2	\$1,663	\$4,427
United States	Men's Bras	Copenhagen	1	\$36,402	\$60,792
United States	Men's Bras	Copenhagen	2	\$72,415	\$26,37
United States	Slippers	Copenhagen	4	\$16,396	\$55
United States	Sport Shoes	Copenhagen	1	\$1,627	\$1
United States	Sport Shoes	Copenhagen	10	\$171,927	\$1

No structure changed in the resulting data.

MODE SELECTION = “translate”

Original
Data

Original **LB** Dataset

SUBJID	LBTESTCD	LBORRES
1001001	A1CHY	16
1001002	A1CHY	25
1001003	A1CHY	18
1001007	A1CHY	4
1001009	A1CHY	12

Original **AE** Dataset

SUBJID	AETERM	AESTDTC
1001001	Broken Crown on...	2013-04-10
1001002	Pyngitis [due t...	2012-11-25
1001003	Pyngitis [due t...	2012-11-25
1001007	mach virus	2012-12-17
1001009	omiting	2012-09-29

Definition
Data set

table	column	mode	identifier_type	date_type	full_date	LABEL
AE	AEDECOD	none				AE Term : MedD...
AE	AEDT	date	date	sasdate		Imputed AE end ...
AE	AEENDTC	date	datec	is8601da.		
AE	AEENDTN	date	datetime	sasdatetime		End Date/Time o...
AE	ORIGSITE	translate				Original Site at S...
AE	SUBJID	translate	group 1			Subject Identifier
LB	SUBJID	translate	group 1			Subject Identifier

Anonymized
Data

Anonymized **LB** Dataset

SUBJID	LBTESTCD	LBORRES
nnnnnn1	APOA1CHY	16
nnnnnn2	APOA1CHY	25
nnnnnn3	APOA1CHY	18
nnnnnn4	APOA1CHY	4
nnnnnn5	APOA1CHY	12

Anonymized **AE** Dataset

SUBJID	AETERM	AESTDTC
nnnnnn1		2418-12-18
nnnnnn2		2418-08-04
nnnnnn3		2418-08-04
nnnnnn4		2418-08-26
nnnnnn5		2418-06-08

Results: Real data Vs. anonymized data

Table 14.3-x.x (Page 1 of 1)
Overall exposure to study drug
Safety set

Statistic	Trt 1 N=338	Trt 2 N=339	Trt 3 N=328
Exposure (in days)			
n	338	339	328
Mean(SD)	322.5 (102.54)	337.2 (87.58)	340.1 (83.51)
Median	365.0	365.0	365.0
Min - Max	5.0 - 407.0	1.0 - 400.0	2.0 - 400.0
Categorized exposure			
<= 4 weeks	n(%) 16 (4.7)	12 (3.5)	12 (3.7)
> 4 - 12 weeks	n(%) 11 (3.3)	8 (2.4)	4 (1.2)
> 12 - 28 weeks	n(%) 19 (5.6)	12 (3.5)	8 (2.4)
> 28 - 40 weeks	n(%) 8 (2.4)	3 (0.9)	5 (1.5)
> 40 - 53 weeks	n(%) 264 (78.1)	286 (84.4)	281 (85.7)
> 53 weeks	n(%) 20 (5.9)	18 (5.3)	18 (5.5)

		trt 1	trt 2	trt 3
N		338.0	339.0	328.0
Exposure (in days)	n	338	339	328
	Mean	322.5	337.2	340.1
	Standard Deviation	102.54	87.58	83.51
	Median	365.00	365.00	365.00
	Min	5.0	1.0	2.0
	Max	407.0	400.0	400.0
Categorized exposure				
<= 4 weeks	n	16.0	12.0	12.0
	%	4.7	3.5	3.7
> 4 - 12 weeks	n	11.0	8.0	4.0
	%	3.3	2.4	1.2
> 12 - 28 weeks	n	19.0	12.0	8.0
	%	5.6	3.5	2.4
> 28 - 40 weeks	n	8.0	3.0	5.0
	%	2.4	0.9	1.5

Conclusion

Programming solution independent of data source structure (CDISC/non CDISC) and SAS versions used

Validated with most stringent validation guidelines to minimize end user QC required

Can handle multiple studies (core/extension) to keep same anonymization parameters

No Key / translation tables retained so link to original data is destroyed

Thank You!

Lets keep sharing ideas and solutions



Acknowledgement

Gargi Paigaonkar

(IQS Global Franchise Head - PC/EM, SR)

Guillaume Breton

(IQS Global Franchise Head - CC, SR)