Patient Confidentiality Implementation

BBS-EFSPI Seminar

November 2014

Guillaume Breton

NOVARTIS

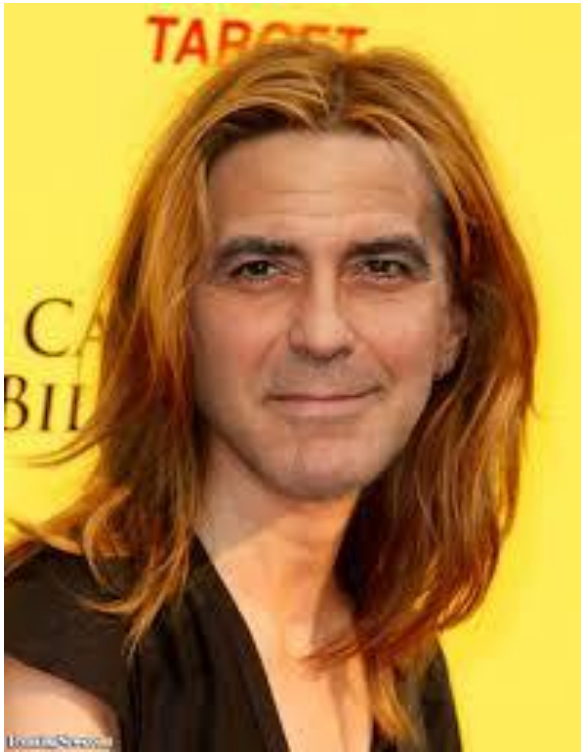
Outline

Topics covered in today presentation

- Defining Industry Standards
- Novartis approach
- Overview of the anonymization process
- Modes of anonymization
- 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

NOVARTIS

Defining Industry Standards

Starting with SDTM

- A PHUSE Working group* has been set up to assess data privacy of all elements of the SDTM Implementation Guide

**led by Jean Marc Ferran (Qualiance) with representatives from the Industry (CRO, Pharma) and Academics*

- Goal is to define if a variable is a **Direct identifier** (a patient can be directly identified) or a **Quasi identifier** (a "merge away" to re-identify the patient) and establish common rules for de-identification (mask, drop, set to missing etc...)

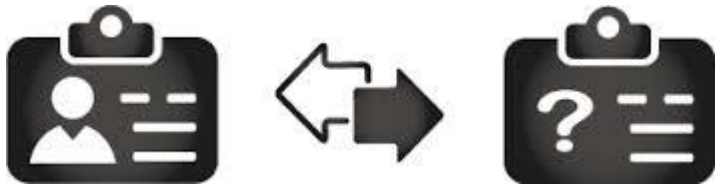
- This initiative among many others, will help to foster de-identification standards and leverage knowledge on data privacy impact and data handling within the industry.

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.

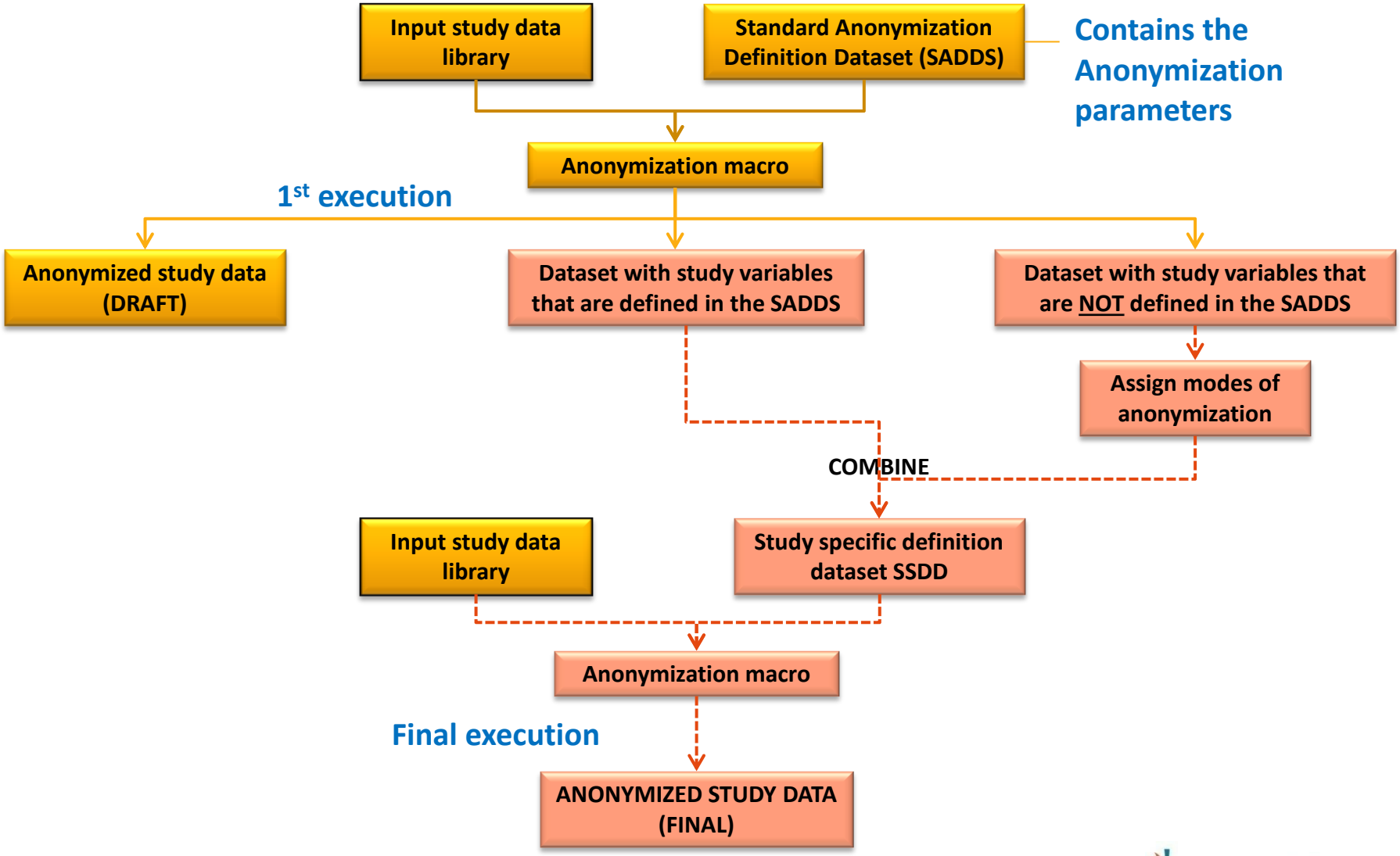


The **Standard** Anonymization Definition Dataset (SADDS) is used across studies and is a repository of standard data definition and mode attributes.

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



NOVARTIS

A High-level view of the Definition Dataset

DATASET	VARIABLE	LABEL	MODE	dentifier_ type	date_type	full_date	TYPE	FORMAT
AEFF2EVT	ACSDT	Date of 1st Hosp. for ACS	DATE	late	sasdate		1	DATE
ACMDATC	CMDEND1O	Concomitant med. end date (Oracle date)	DATE	latetime	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)

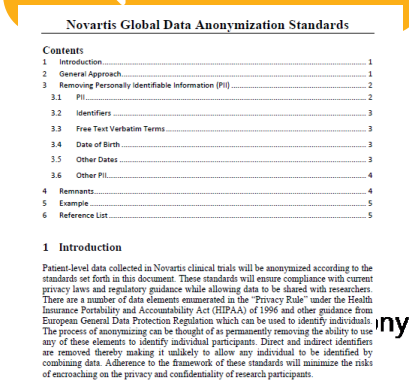
Process flow

Novartis Anonymization Guidelines

Anonymization Mode Selection

Programming Considerations

Output Anonymized data



Novartis Global Data Anonymization Standards

Contents

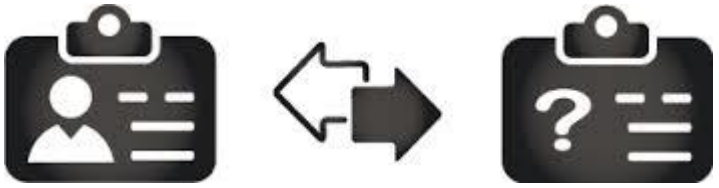
1 Introduction

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

Upon approved requests, the following data and accompanying trial documentation will

Novartis
nymization Guidel



NOVARTIS

Modes of anonymization

Walkthrough and examples

Following are the modes of Anonymization:

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

MODE = “none”

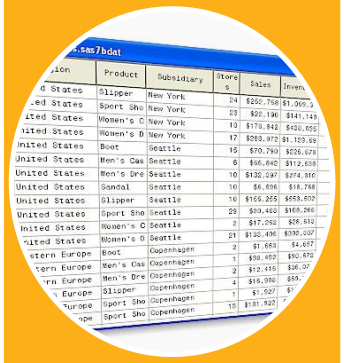
Straight copy of the variable

3.3.2. All standard dictionary coded terms will be retained.

AEDECOD will not be anonymized.

MODE = “none”

Variable values will be unchanged.



No structure changed in the resulting data.

NOVARTIS

MODE SELECTION = “none”

Original Data

SUBJID	AETERM	AESTDTC	AEDECOD	AEENDTN	INV	ORIGSITE
1001001	Broken Crown on...	2013-04-1	DEVICE BREAK...	13-04-10T00:0...	308667C@CA	1011
1001002	pharyngitis [due t...	2012-11-	PHARYNGITIS	2-12-02T00:0...	308679C@CA	1017
1001003	pharyngitis [due t...	2012-11	PHARYNGITIS	2-12-02T00:0...	308679C@CA	1017
1001007	stomach virus	2012-12-	GASTROENTER	2-12-18T00:0...	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	AEENDTC	date	datec	is8601da.		
AE	A		datetime	sasdatetime		End Date/Time o...
AE	AETE	NDTN	datec	is8601da.		Start Date/Time ...
AE		dat				Reported Term fo...

Anonymized Data

SUBJID	AETERM	AESTDTC	AEDECOD	DTN	INV	ORIGSITE
nnnnnn1		2323-07-1	DEVICE BREAK...	18T00:0...	nnnnnnX@11	nnn5
nnnnnn2		2323-03-	PHARYNGITIS	1T00:0...	nnnnnnX@12	nnn6
nnnnnn3		2323-03-	PHARYNGITIS	1T00:0...	nnnnnnX@12	nnn6
nnnnnn4		2323-03-2	GASTROENTER	7T00:0...	nnnnnnX@X4	nnnB
nnnnnn5		2323-01-06	VOMITING	-10T00:0...	nnnnnnX@XD	nnnC

MODE = “missing”

Keeping variable but removing its contents

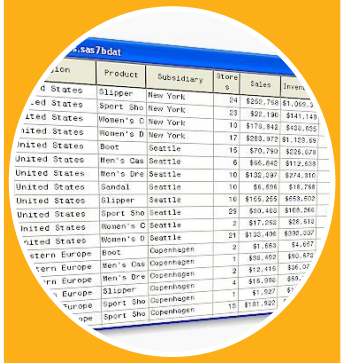
3.3.1. Free text verbatim terms are not included and are set to blank or dropped from the following datasets:

- 3.3.1.1 Adverse Events
- 3.3.1.2 Medications
- 3.3.1.3 Medical History
- 3.3.1.4 Other specific verbatim free text

AETERM will be set to blank.

MODE = “missing”

Variable values will be missing in the anonymized dataset.



No structure changed in the resulting data.

NOVARTIS

MODE SELECTION = “missing”

Original Data

SUBJID	AETERM	AESTDTC	AEDECOD	AEENDTN	INV	ORIGSITE
100100	Broken Crown on...	013-04-10	DEVICE BREAK...	2013-04-10T00:0...	308667C@CA	1011
10010	pharyngitis [due t...	12-11-25	PHARYNGITIS	2012-12-02T00:0...	308679C@CA	1017
10010	pharyngitis [due t...	12-11-25	PHARYNGITIS	2012-12-02T00:0...	308679C@CA	1017
100100	stomach virus	012-12-17	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	conn	mode ode	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						

Anonymized Data

SUBJID	AETERM	AE TC	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			PHARYNGITIS	2323-03-11T00:0...	nnnnnnX@12	nnn6
nnnnnn4		2323-0	GASTROENTER...	2323-03-27T00:0...	nnnnnnX@X4	nnnB
nnnnnn5		232 06	VOMITING	2323-01-10T00:0...	nnnnnnX@XD	nnnC

MODE = “drop”

Dropping variable

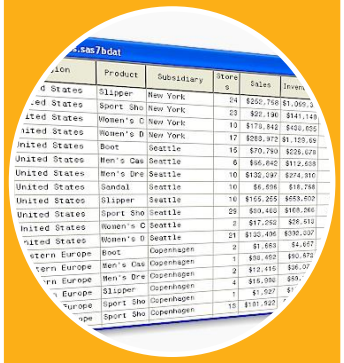
3.4 Date of Birth

- 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.



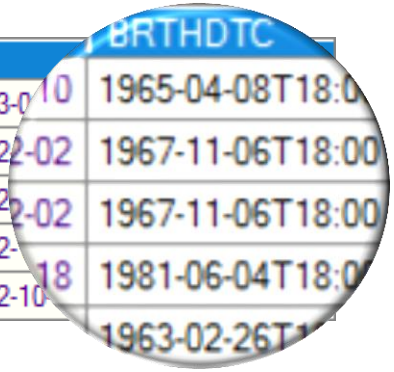
Structure will be changed by dropping a column in the resulting data.

NOVARTIS

MODE SELECTION = “drop”

Original Data

ORIGSITE	AEENDTC	AGE	AEEDT
1011	2013-04-10T00:0...	47	2013-0
1017	2012-12-02T00:0...	45	2012-02
1017	2012-12-02T00:0...	97	2012-02
1023	2012-12-18T00:0...	31	2012-
1025	2012-10-03T00:0...	49	2012-10



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	BRTHDTC	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)*

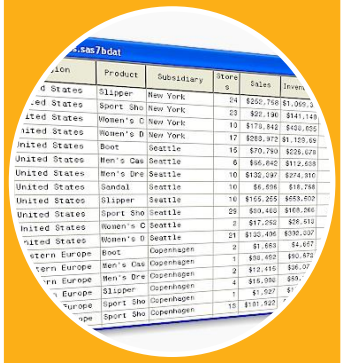
3.4 Date of Birth

- 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 aggre ed into a sing or

AGE above 89 will be grouped into a category of ‘90 or older’, the rest will be output as is.

MODE = “ageint”

For numeric variable, create a character variable with the category.



If a numeric age variable exists, then a new character variable named *<Original variable name>*_cat will be created and the original numeric variable will be dropped.

**US Health Insurance Portability and Accountability Act (1996)*

NOVARTIS

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	BRTHDTC	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

3.5 Other Dates Specific dates directly related to a research participant may compromise a research participant’s anonymity.

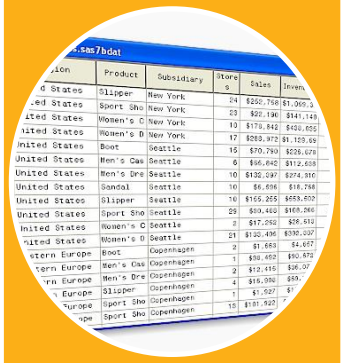
- 3.5.1. A random offset per study, is generated and applied to all dates. All original dates are replaced with the new dummy dates so that the relative times between dates are retained.
- 3.5.2. This date offset will be generated at the study level pushing dates into the future.

AEENDTC will be offset to a future date.

MODE = “date”

The intervals between dates in anonymized data must be consistent with those in original data.

All dates will be offset by a random number between 200 years and 100,000 days after that.



No structure changed in the resulting data.

NOVARTIS

MODE SELECTION = “date”

Original Data

INV	ORIGSITE	AEENDTC	AGE	AEEDT	
308667C@CA	1011	2013-04-10T00:0...	47	2013-04-10	19
308679C@CA	1017	2012-12-02T00:0...	45	2012-12-02	196
308679C@CA	1017	2012-12-02T00:0...	97	2012-12-02	196
127996C@CA	1023	2012-12-18T00:0...	31	2012-12-18	19
277526C@CA	1025	2012-10-03T00:0...	49	2012-10-03	

Definition Data set

table	column	mde	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 in AGEU at ...

Anonymized Data

AEENDTC	AEEDT	BRTHDTC	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

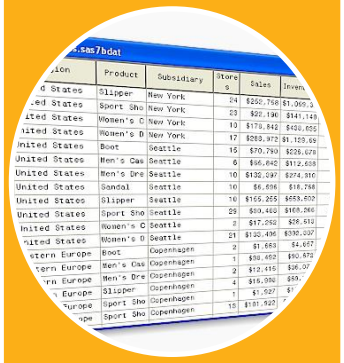
3.2 Identifiers Change the real value to a de-identified value in a consistent manner so that the value in one instance of the variable is consistent with the value in the same variable across other datasets.

- 3.2.1. The investigator number is re-coded or set to blank for each investigator. The investigator name is set to blank or dropped from the data
- 3.2.2 [illegible] 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.



No structure changed in the resulting data.

NOVARTIS

MODE SELECTION = "translate"

Original Data

Original **LB** Dataset

SUBJID	BTESTCD	LBORRES
1001001	A1CHY	16
1001002	1CHY	25
1001003	CHY	18
1001007	ICHY	4
1001009	A1CHY	12

Original **AE** Dataset

SUBJID	AETERM	AESTDTC
1001001	oken Crown on...	2013-04-10
1001002	yngitis [due t...	2012-11-25
1001003	yngitis [due t...	2012-11-25
1001007	nach 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	AEEDT	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	group1			Subject Identifier
LB	SUBJID	translate	group1			Subject Identifier

Anonymized Data

Anonymized **LB** Dataset

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

Anonymized **AE** Dataset

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

Conclusion

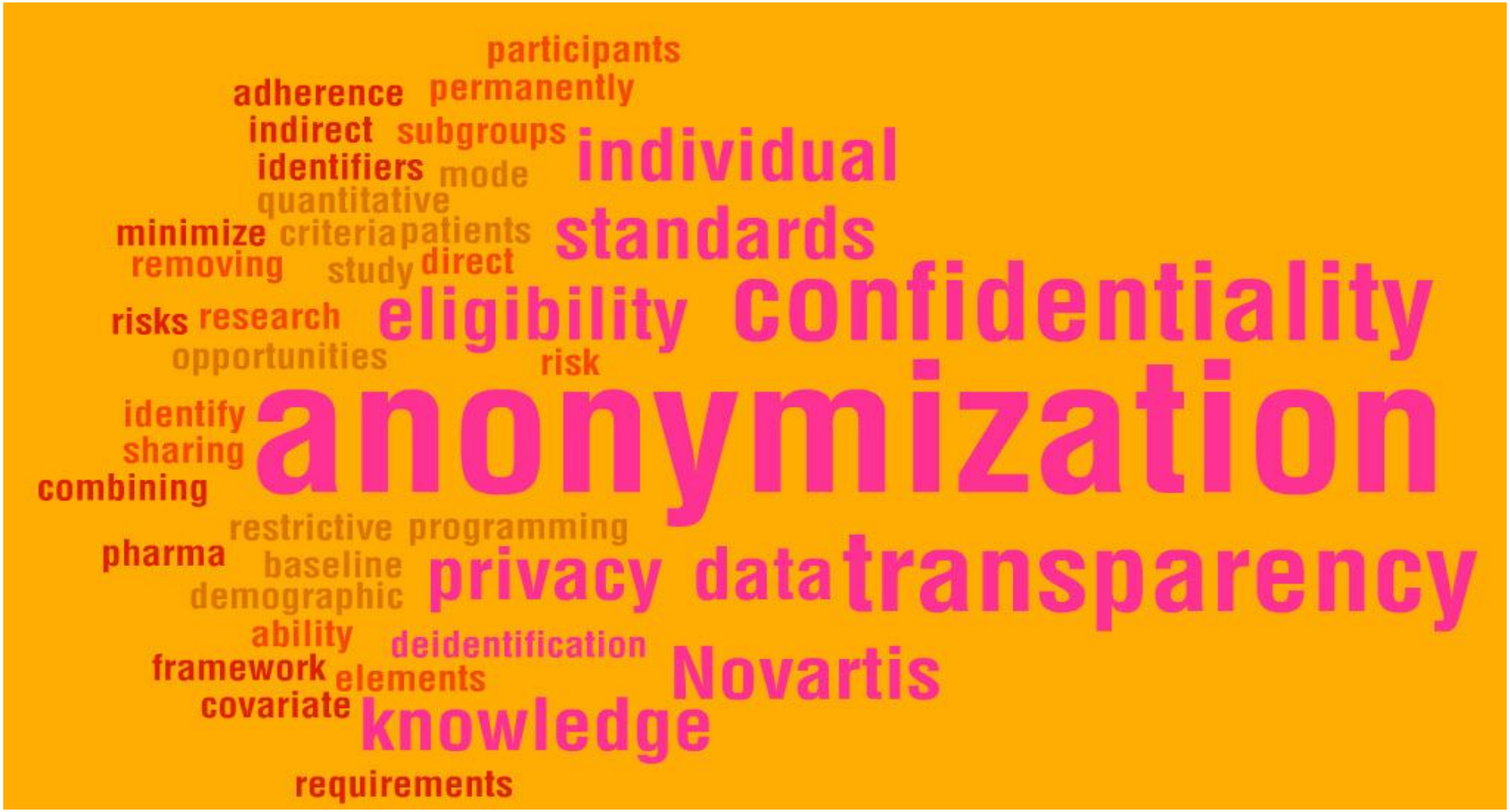
Programming solution independent of data source structure (CDISC/non CDISC) and SAS versions used

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 - QA



NOVARTIS