PhUSE 2016

Paper PP22

KeyDAN: Risk Minimization in Data Anonymization

Stéphane Chollet, Keyrus Biopharma, Levallois-Perret, France
Mathilde Laffitte, Keyrus Biopharma, Lasne, Belgium

ABSTRACT

Anonymizing clinical data to allow research while maintaining patient privacy and confidentiality is becoming mandatory and remains a challenge.

As a CRO, Keyrus Biopharma needs to provide the sponsor with an efficient anonymization solution, combining in-house anonymization software, internal expertise and robust processes with final QC and re-identification risk assessment. KeyDAN tool allows anonymization of data whatever the database structure, with an easy-to-use, efficient tool which can anonymize data within any database structure.

Date anonymization is one of the most important challenges. In order to facilitate the identification of variables which are dates or contain a date (in character format) we propose an add-on which determines the probability to be a date whatever the format (numeric or character).

After anonymization, another challenge is to identify whether some variables, which have been left un-anonymized, are derived or contain information from variables that have been anonymized. The key item detector add-on answers to this challenge.

This article presents the data anonymization process as well as anonymization examples together with the algorithm of automatic date detection.

INTRODUCTION

Anonymizing clinical data to ensure patient privacy and confidentiality remains a challenge, especially for **non-standard** or **legacy** structures (not CDISC like) and **SUPPQUAL CDISC** tables.

To aid the anonymization process, Keyrus Biopharma (KBP) has developed an easy-to-use, efficient tool, named KeyDAN, which can anonymize data within any database structure.

As with any other anonymization solutions the two most important challenges for data managers are:

- Detecting all the variables to be anonymized;
- Associating the right anonymization functions.

KBP's tool scans the entire database to automatically detect:

- The variables that can be considered as dates;
- Non-anonymized variables that contain, as a substring, key identifiers which must be anonymized.

This article presents the two solutions that we have implemented to **limit the risk of missing potential re-identifying information.**

DATA ANONYMIZATION PROCESS & KBP ANONYMIZATION TOOL

Based on data privacy and data protection regulations/directives, pharmaceutical companies and biotechs published a process and detailed anonymization rules to apply before clinical data sharing. These were summarized in the document 'Model Approach: De-identification and Anonymization of Privacy Information in Data – © 2013 TransCelerate BioPharma Inc.'.

TransCelerate BioPharma Inc. is a non-profit organization with a mission to collaborate across the biopharmaceutical research and development community to identify, prioritize, design and facilitate the implementation of solutions to drive efficient, effective and high-quality delivery of new medicines, improving the health of people around the world.

The clinical data sharing objective is to increase transparency of clinical results:

- Past: posting of clinical report summaries;
- Currently: data sharing of clinical results with external researchers, in addition to clinical report summaries.
- Future: sharing anonymized Clinical Reports and Clinical Data as per EMA Policy 0070.

The clinical data sharing process can be summarized as follows:

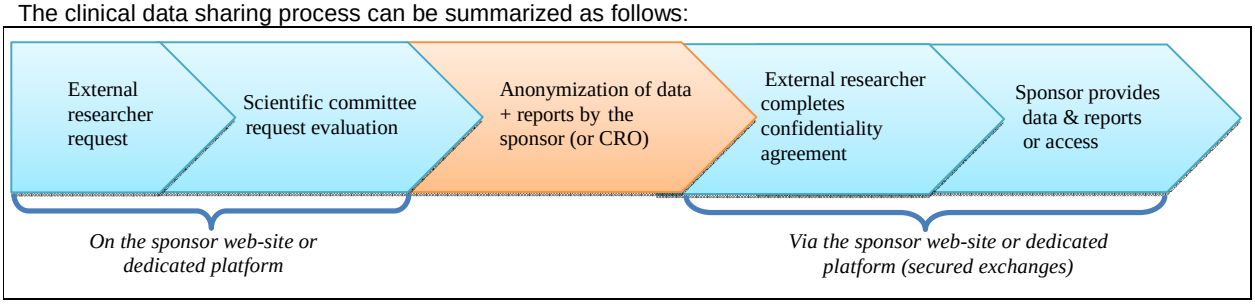


Figure 1: Clinical data sharing process

KBP has developed KeyDAN, a SAS application to help the data manager to target the anonymization of clinical data (Figure 1 - step 'Sponsor Anonymizes Data'). The tool complies with the following requirements:

- Anonymize data consistently across all source datasets for each subject and provide final anonymized datasets;
- Reproduce the basic anonymization functions and algorithms as described in the TransCelerate document (summarized in Table 1);
- Provide the data manager a functionality to scan the source data in order to facilitate identification of variables to be anonymized;
- Allow data manager to allocate anonymization functions coming from standard structure (e.g. PhUSE De-identification Standard for SDTM 3.2, sponsor standard…) and from clinical data previously anonymized;
- Allow anonymization functions allocation whatever the database structure (horizontal or vertical structure like SUPPQUAL CDISC-SDTM datasets).

Function	Action
Variable identifier	Provide a new identifier.
Subject identifier	Provide a new identifier and identifies the datasets where functions Date, Date of birth and Age can be allocated.
Date	*Offset Date*: Provide modified date by adding or removing random days for the subject (consistently across all dates for the subject) and repeat the operation for each subject. *Relative Study Day*: Delete date and replace by relative study day for the subject defining an anchor date (variable) (i.e. informed consent date, 1st visit date, …)
Date of Birth	Apply *Offset Date* calculation and remove day and month for subjects ≤ 89 Years – delete date of birth for subjects > 89 Years – or delete all dates of birth.
Age	Delete all values and create a new variable with age categories or keep values for age and delete the ones for the subjects older than 89 years and create two or more age-categories (i.e. > 89 and ≤ 89)
Set variable to blank	Keep the variable and delete contents for all records
Drop the variable	The variable is removed from the database

Table 1: Anonymization functions

In order to provide anonymized clinical data, equivalent to the original data, the tool is used in 3 steps, corresponding to the 3 main functionalities and deliverables:

1. Plan builder: generates the list of tables and variables from clinical data to be anonymized. Based on this list, the data manager has to allocate anonymization function on appropriate variables.

2. Anonymization tool: anonymization is conducted according to the plan. The module anonymizes the database, including normalized datasets (vertical structure), based on input parameters, allocated anonymization functions and standard anonymization functions allocation (if applicable).

3. Report builder: the report builder generates an automatic QC report and assessment report.

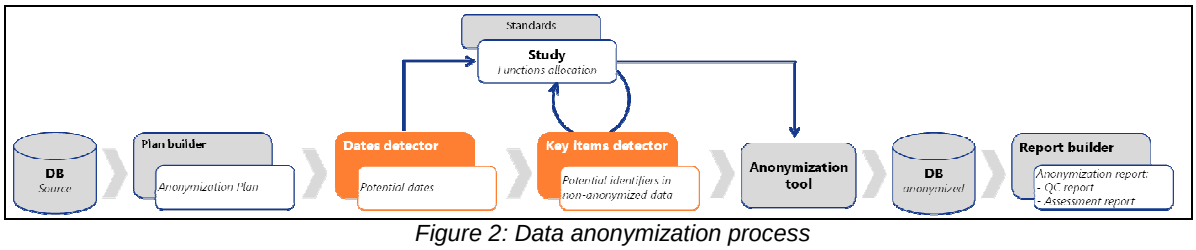


Figure 2: Data anonymization process

Once the process and tools are in place, anonymization functions must be manually allocated to the variables. Misallocations may introduce a risk of missing potential identifiable information. Among all potential re-identification risks, we have developed add-on tools to focus on two specific elements:

- Identification of items that can be considered/interpreted as a date (*dates detector* add-on);
- Identification of items (without anonymization function allocated) that can contain information coming from original values of variables with anonymization function allocated (*key items detector* add-on).

DATES DETECTOR

For numeric variables, the date detection is easy using the SAS format/informat (DATE9., DDMMYYY., B8601DA8., …) metadata coming from *Sashelp.vcolumn* table.

Character variable can contain date information (e.g. “03OCT2016”, “03/10/2016” …) or not (e.g. “FEVER”, “250MG” …). In addition, a same character variable can even contain a mix of information (dates or not) and, if containing only dates, a mix of date formats. In such cases, the detection becomes more difficult because information stored in *Sashelp.vcolumn* is no longer sufficient.

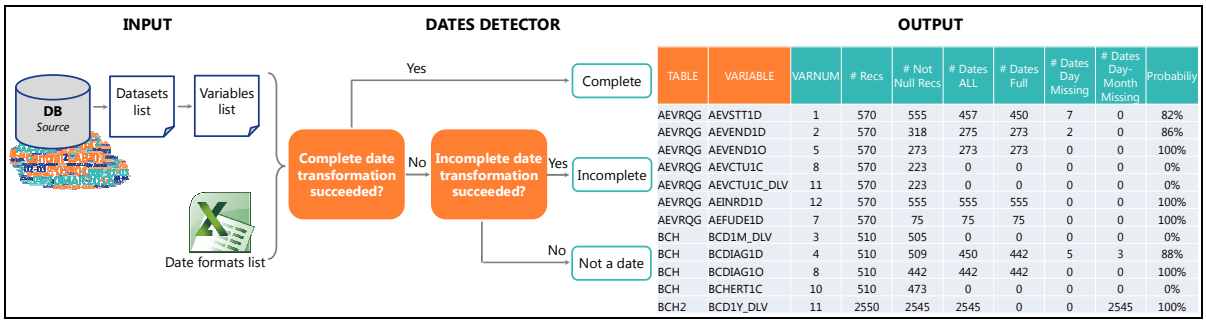


TABLE	VARIABLE	VARNUM	# Recs	# Not Null Recs	# Dates ALL	# Dates Full	# Dates Day Missing	# Dates Day-Month Missing	Probability
AEVRQG	AEVSTT1D	1	570	555	457	450	7	0	82%
AEVRQG	AEVEND1D	2	570	318	275	273	2	0	86%
AEVRQG	AEVEND1O	5	570	273	273	273	0	0	100%
AEVRQG	AEVCTU1C	8	570	223	0	0	0	0	0%
AEVRQG	AEVCTU1C_DLV	11	570	223	0	0	0	0	0%
AEVRQG	AEINRD1D	12	570	555	555	555	0	0	100%
AEVRQG	AEFUDE1D	7	570	75	75	75	0	0	100%
BCH	BCD1M_DLV	3	510	505	0	0	0	0	0%
BCH	BCDIAG1D	4	510	509	450	442	5	3	88%
BCH	BCDIAG1O	8	510	442	442	442	0	0	100%
BCH	BCHERT1C	10	510	473	0	0	0	0	0%
BCH2	BCD1Y_DLV	11	2550	2545	2545	0	0	2545	100%

Figure 3: Date detection algorithm

In all CDISC domains, date variables are fixed by the SDTMIG documentation. A data manager can therefore immediately identify which variables have to receive date anonymization functions.

However, in the CDISC SUPPQUAL domains, all clinical data information is stored into one variable (QVAL), which can contain a mix of different information e.g. laboratory name, date of sampling, sampling code etc. It is important to determine if such variables are date free or not, so that, in a second step, the records which are dates for this variable can be anonymized consistently.

With legacy databases, the information on variable type and variable descriptions is generally available in the study documentation (provided by the sponsor). However, identifying variables that are dates without missing any, especially for the variables with missing data, is time consuming and mistakes are likely.

TABLE	VARIABLE	VARNUM	# Recs	# Not Null Recs	# Dates ALL	# Dates Full	# Dates Day Missing	# Dates Day-Month Missing	DATE7 Char_full	DATE7 Char_miss_ days	DATE7 Char_miss_ dmth	DATE9 Char_full	DATE9 Char_miss_ days	DATE9 Char_miss_ dmth	ISO_DATE 8 Char_full	ISO_DATE8 Char_miss_d ays	ISO_DATE8 Char_miss_ dmth
AEVRQG	AEVSTT1D	30	570	555	457	450	7	0	0	1	0	450	6	0	0	0	0
AEVRQG	AEVEND1D	41	570	318	275	273	2	0	0	0	0	273	2	0	0	0	0
AEVRQG	AEVEND1O	42	570	273	273	273	0	0	0	0	0	0	0	0	273	0	0
AEVRQG	AEVCTU1C	44	570	223	0	0	0	0	0	0	0	0	0	0	0	0	0
AEVRQG	AEVCTU1C_DLV	45	570	223	0	0	0	0	0	0	0	0	0	0	0	0	0
AEVRQG	AEINRD1D	55	570	555	555	555	0	0	0	0	0	0	0	0	555	0	0
AEVRQG	AEFUDE1D	58	570	75	75	75	0	0	0	0	0	0	0	0	75	0	0
BCH	BCD1M_DLV	15	510	505	0	0	0	0	0	0	0	0	0	0	0	0	0
BCH	BCDIAG1D	19	510	509	450	442	5	3	0	0	3	442	5	0	0	0	0
BCH	BCDIAG1O	20	510	442	442	442	0	0	0	0	0	0	0	0	442	0	0
BCH	BCHERT1C	31	510	473	0	0	0	0	0	0	0	0	0	0	0	0	0
BCH2	BCD1Y_DLV	18	2550	2545	2545	0	0	2545	0	0	2545	0	0	0	0	0	0

Table 2: Date detector detailed output

Within multiple data sources (e.g. raw and analysis datasets) based on a pre-defined date format list, the *date detector* add-on scans all the database tables and variables and checks whether values could be interpreted as a date. A report by table-variable is generated with:

- Number of records;
- Number of non-missing records;
- Numbers of records that are detected as a date (complete or incomplete);
- Numbers of records by date formats.

Missing day-month symbols (UNK, NK …) can also be used as an input parameter.

The KeyDAN *date detector* add-on information can be used to compute the number of records successfully considered as a date (among the list of formats provided as input) as described in the above table, allowing the data manager to investigate and determine a "*where clause*" (selection of items inside a same variable answering a specific condition for vertical datasets structure like SUPPQAL CIDSC datasets) when associating the anonymization function.

For example we will use QNAM = "COLDATE2" as a "*where clause*" to select all rows with "COLDATE2" as value of the QNAM variable in a SUPPQUAL table. The purpose is to retrieve rows where QVAL value needs to be anonymized.

KEY ITEMS DETECTOR

The detection of variables that can be interpreted as a date was our first milestone in reducing re-identification risks during data anonymization.

Our second milestone was the detection of identifying information. This information is stored in the values of dedicated variables (e.g. subject identifiers, center numbers) but also as part of generic variables content (e.g. surgical procedure date present in a SAE verbatim, laboratory identifier in a reason for sample not done).

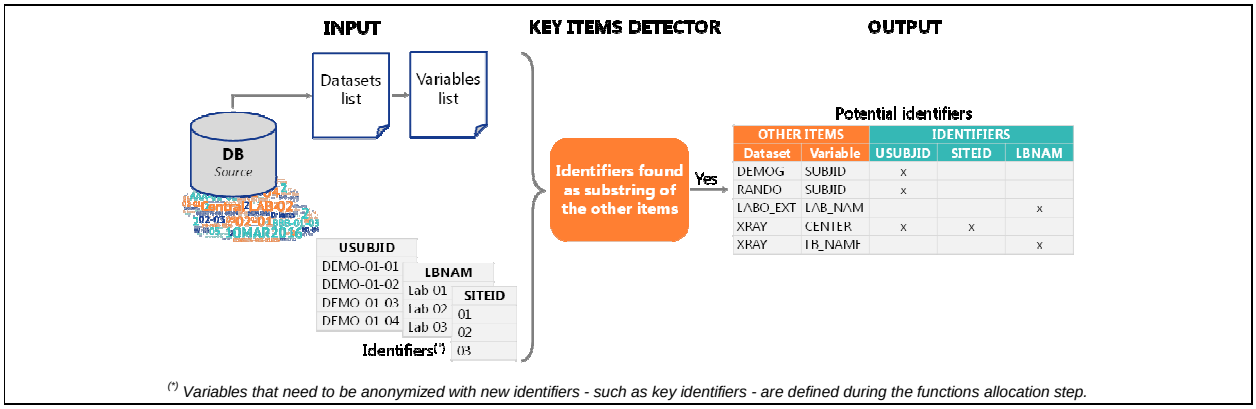
At time of the anonymization, the list of variables is screened by the data manager who decides whether an anonymization function has to be allocated or not. This decision is described in the anonymization plan, and is driven by the study documentation, mainly the database documentation and annotated CRF.

To reduce the risk of missing some data to be anonymized, we developed the *key item detector* add-on tool to check that variables – for which no anonymization functions has been allocated – do not contain information coming from variables that have to be anonymized.

For example, in the following comment field "*Pr. Martin ensures that the result is not clinical significant regarding patient's medical history*", the investigator's name is mentioned. If no anonymization function is allocated on the variable corresponding to this comment, the name of the investigator will not be removed and is at risk for re-identification.

Then, consider that the variable containing the investigator's name ("*Pr. Martin*") is targeted as a variable to be anonymized by the data manager.

It is critical to detect that the comment field contains information that must be anonymized. That is what the KeyDAN *key item detector* add-on does.



(*) *Variables that need to be anonymized with new identifiers - such as key identifiers - are defined during the functions allocation step.*

Figure 4: Key item detector process

Missing cases such as this whilst anonymizing a database is not acceptable. Checking the information depends on the allocated anonymization function. The complexity and the hardware resources needed for program execution can vary significantly from several minutes to several hours. But we can sort the anonymization functions into 3 main classes:

Class 1 for Subject ***Identifier*** and ***Variable Identifier*** functions:

Distinct values can easily be retrieved from identifiers defined during the function allocation step. A limited set of different values is expected, allowing a good execution time.

Class 2 for ***Drop*** and ***Blank*** functions:

Looking for the exact match of the content in other items as substring and replacing the value by a set of XXXX value leads to a longer execution time due to the larger set of distinct values. Nevertheless it is still acceptable when the SAS program is launched in batch mode or on a dedicated server.

Class 3 for ***Dates***, ***Date of Birth*** and ***Age*** functions:

These functions are more complicated because a data or an age can be expressed in several formats, not necessarily the one used for the dedicated item.

The current version of KeyDAN integrates the class 1 functions (i.e. functions ***Subject Identifier*** and ***Variable Identifier***) and the next versions will integrate the two remaining function classes.

The *key items detector* add-on scans all records in the database for variables (character or numeric) which are not anonymized and provide a report with all the non-anonymized data that may contain potential identifying information. If any record looks like a key identifier value or contains the value as a substring, then the corresponding variable is highlighted. Those records are at risk for subject re-identification and two actions are possible:

- Update the content of these variables in replacing the identifiable information by the anonymized one;
- Allocate an anonymization function to these variables.

The *key item detector* add-on reduces the risk of missing variables requiring anonymization. It can also be used as a quality control tool at the end of the anonymization process.

CONCLUSION

Reducing the risk of missing potential identifiable information in data which will be made available to external users – to maintain transparency and provide exploitable data – is a difficult and time-consuming obligation to comply with.

The KeyDAN add-ons we have developed help the data manager at the beginning of the anonymization process to identify:

- Which variables are at risk of being missed as a date which should be anonymized;
- Which variables are at risk of containing a value coming from another anonymized variable.

The *date detector* and *key item detector* add-ons significantly decrease data manager working time and increases quality by minimizing the risks in the data anonymization process.

REFERENCES

Code of Federal Regulations - Title 45: Public Welfare, Subtitle A §164.514: Available at http://www.ecfr.gov/cgi-bin/text-idx?tpl=/ecfrbrowse/Title45/45cfr164_main_02.tpl

Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule – http://privacyruleandresearch.nih.gov/pr_08.asp#8a

Council Regulation (EC) 45/2001 of the European Parliament and of the Council of 18 Dec 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data [2008] OJ L193/7.

Council Directive (EC) 95/46 of the European parliament and of the council of 24 Oct 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data [1995] OJ L281.

Regulation (EU) 536/2014 of the European parliament and of the council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC [2014] OJ L158/1.

European Medicines Agency: EMA/240810/2013 - Publication of clinical data for medicinal products for human use. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf

Article 29 Data Protection Working Party: 0829/14/EN WP216 - Opinion 05/2014 on Anonymisation Techniques. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

European Medicines Agency policy on publication of clinical data for medicinal products for human use: EMA/240810/2013– 02Oct2014 http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf

Data De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach - 2013 TransCelerate BioPharma Inc. All rights reserved – http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/CDT-Data-Anonymization-Paper-FINAL.pdf

ACKNOWLEDGMENTS

Cathy Scoupe, Head operations Belgium, Keyrus Biopharma
Emmeline Jallas, Marketing specialist, Keyrus Biopharma
Jean Fernandez, Biometry expert, Keyrus Biopharma (JFE Consulting)
Kirsten Dumaz, Proposal manager, Keyrus Biopharma

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Author Name: Stéphane Chollet
Company: Keyrus Biopharma
Address: 18-20, rue Clément Bayard
City / Postcode: F-92300 Levallois-Perret - France
Email: stephane.chollet@keyrus.com
Web: http://www.keyrusbiopharma.com/

Author Name: Mathilde Laffitte
Company: Keyrus Biopharma
Address: Chaussée de Louvain, 88
City / Postcode: B-1380 Lasne - Belgium
Email: mathilde.laffitte@keyrus.com
Web: http://www.keyrusbiopharma.com/