PhUSE 2017

Paper SD01

KeyDAN, a Clinical Data De-identification Solution

Stéphane Chollet, Keyrus Biopharma, Levallois-Perret, France

ABSTRACT

The clinical data anonymization to allow research while maintaining subject privacy and confidentiality is becoming mandatory. As a CRO, Keyrus Biopharma has developed KeyDAN tool to provide the sponsors an efficient anonymization solution.
KeyDAN is structured as an end-to-end solution for pseudonymization or anonymization. KeyDAN offers a wide set of functionalities helping in processing the data by:

- Managing dates whatever their completeness and format.
- Detecting all the anonymized variables present, as part (or not) of non-anonymized ones.
- Assessing the n-uplet combination frequency of variables identified as being of special interest.

These functionalities allow a systematic check of the database to ensure that no critical data were missed and to provide evidence that anonymization minimizes the risk of patient re-identification, in accordance with the new EMA regulations.

During the demo we will present the use of the different add-ons as well as the generated outputs.

INTRODUCTION

Following the implementation of EMA Policy 0070, Applicants/MAHs are required to publish their clinical data by providing the clinical reports (clinical overviews -submitted in module 2.5-, clinical summaries -submitted in module 2.7- and the clinical study reports -submitted in module 5, "CSR"-) as well as the individual subject data (IPD). In parallel, the European Global Data Protection Regulation (GDPR), which will come into force by May 2018 and applies to all data processing conducted in connection with clinical trials, requires that personal data be protected. This presents a challenge to the applicant/MAH: how do you protect the privacy of personal data that are being released to the public domain?

Keyrus Biopharma has developed the KeyDAN tool and add-ons to answer this question.

PSEUDOMINIZATION THEORETICAL FEATURES

CLINICAL DATA DE-IDENTIFICATION PURPOSE

The data variables contained in a clinical database can be divided in 3 main types based on how much information they provide about an individual:

- Direct-identifiers = data variables such as names, addresses or government identity numbers that directly identify an individual
- Quasi-identifiers = data variables that when combined with others allow an individual to be identified
- Generic information = variables that can be shared without causing any possibility of identification.

It is easy to understand that the direct identifiers must to be removed from the database before it is released to the public domain - it does not make sense to put credit card or the cell phone numbers on the Internet, without any restrictions on their accessibility. Generic information (for example the presence of the investigator signature on the CRF, coding dictionary version …) can be shared without causing any risk to the subject privacy. Quasi-identifiers are more complex to manage.

As mentioned, a quasi-identifier is defined as a variable or a combination of variables that can be used indirectly to re-identify individuals (e.g. BMI value, geographical origin, household composition …). Unlike direct-identifiers (e.g. name, e-mail address, social security number …) quasi-identifiers need be linked to external information to re-identify individual personal information.

Disclosure of personal health information could cause severe harm to subjects. It violates not only the subjects' rights to privacy, but also could potentially compromise their economic situation for example by leading to significant

increases in their health insurance contributions, or to the rejections of bank loan or job applications. In order to guarantee the privacy of personal health data, direct identifiers are already removed from clinical databases. The goal of the European GDPR is to ensure that the anonymization of the clinical database is sufficient to prevent the re-identification of subjects even when the data in the database are combined with data from external sources such as published obituaries, Google search results, or external databases (reached legally or not).

RISK ASSESSMENT

To assess the re-identification risk, all publications mentioned 2 mains axis:

- Record uniqueness = how singular is the record for a given subject compared to those of other subjects?
- Threat = how / when / where and by whom could hacking be attempted.

Based on the literature[1, 2], we define the record frequency $\mathbf{f_i}$ to determine the occurrence numbers for each selected variable or combination of variables. It is also possible to compute the probability of record re-identification $\mathbf{_{re\text{-}id}P_i = 1 / f_i}$. Considering that the re-identification of at least one subject is easier than the re-identification of a specific subject, the overall risk for re-identification is $\mathbf{_{re\text{-}id}R = max(_{re\text{-}id}P_i)}$.

All work to de-identify data focuses on transforming the datasets with a set of actions in order to decrease the $\mathbf{_{re\text{-}id}R}$ to an acceptable rate (the threshold determination will be detailed later in paragraph '*Estimating the re-identification risk*'). Based on the TransCelerate model approach[3], the KeyDAN solution proposes the following set of actions for the data transformation:

- Recode identifiers,
- Date translation,
- Delete the information,
- Remove the information.

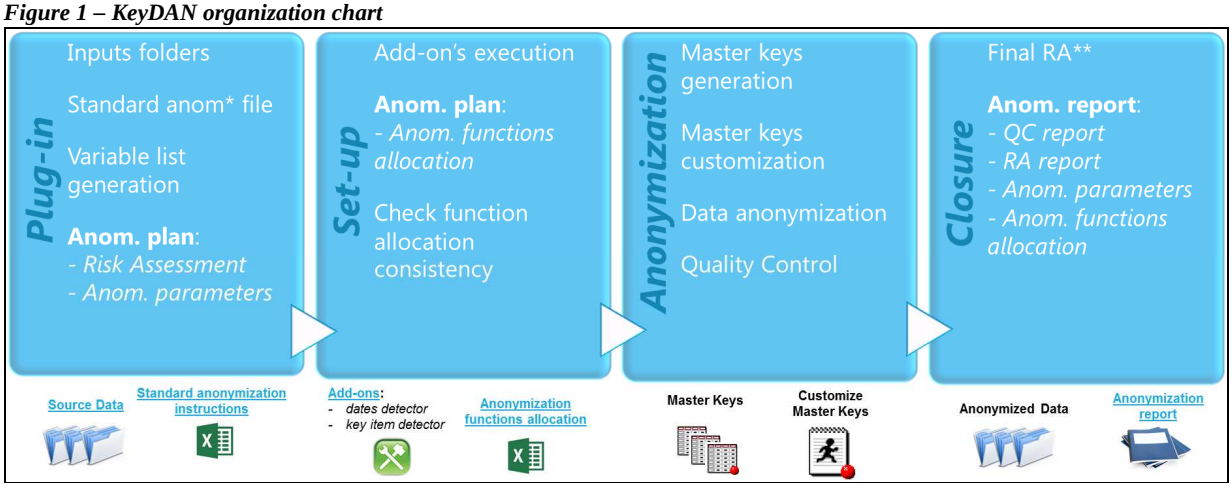
KEYDAN MAIN FEATURES

The KeyDAN solution is composed of integrated SAS macros that allow control of the whole process. These macros use the anonymization plan as input to allow the data de-identification process to be incorporated within the normal data processing system. This avoids the need to transfer information from one system to another, thereby reducing the data QC and potential errors that are associated with such transfers.

As shown in the *Figure 1*, the process is divided into 4 sub-steps that allow progressive execution and corrective actions implementation:

- Configuration: definition of input dataset and standard template (if applicable) locations
- Set-up: add-on execution to help determine the appropriate data transformation
- Anonymization: execution of the data transformation as specified in the parameter's file
- Closure: quality control and reporting

Figure 1 – KeyDAN organization chart



** Anom: anonymization*

*** RA: Risk Assessment*

In addition to the main de-identification process, the KeyDAN solution includes add-ons to perform automatic checks and computations for the assessment and limitation of the re-identification risk. These add-ons are as follows:

- Date detector[4]: detect all the strings that can be interpreted as dates (partial or not).
- Key item detector[4]: detect all records containing as substring the initial value of a variable that will be transformed
- Low freq detector: is used to compute the variable (or combination of variable) number of records per subject (i.e. $\mathbf{f_i}$ values).

ESTIMATING THE RE-IDENTIFICATION RISK

FREQUENCIES GENERATION

As databases may contain hundreds to thousands of items, the combination of variables is huge and cannot be determined using normal SAS servers. Moreover, trying the combination of all variables will create false positive signals because with a certain number of variables, all the records will be unique. That's why we ask the user to use the list of tables / variables (for the anonymization function allocation) to flag items of special interest based on the pathology, the primary objective of the study, and so on.

Based on this list, the program generates all combinations from 1 to N variables of 1Cn *(with N the number of flagged variable, and n all the integer values between 1 and N)* and computes the numbers of subjects per distinct value of the combination. The output is sorted by frequency ascending values and is used to compute the maximal risk for a subject to be re-identified[1] as $\mathbf{max(_{re\text{-}id}P) = 1 / min(f_i)}$.

THREAT MODELING

Depending on the data addressee, literature[1, 2, 3] provides attack probabilities based on the data location and the motivation of the aggressor to re-identify the data. Using such modeling will allow data to be handled in a secure location more conservatively than in the public domain.

THRESHOLD ASSESSMENT

The computation of indicators to measure the risk level associated with a future treatment leads to a threshold definition before the processing. Depending on the parameters provided in the clinical study report (as average, standard deviation …) the de-identification process should introduce enough uncertainty to prevent missing value computation.

For example, applying the basic rules for anonymization to the age dataset {70; 72; 75; 77; 80; 82; 85; 87; 90}, leads to the transformed dataset {70; 72; 75; 77; 80; 82; 85; 87; >89}. If the average value is presented in the CSR, it can be used to easily retrieve the original age value '90' from the transformed value '>89'.

That's why (in a first approach) we will fix the minimum number of subjects per value (**x**) to the number of parameters provided in the CSR +1 (in a future publication we will present the difference case scenarios for the threshold determination). The goal of the de-identification process will then be to target a $\mathbf{max(_{re\text{-}id}P) \leq 1 / x}$.

CONCLUSION

The KeyDAN solution has been developed to automate the de-identification process whatever the data structure (legacy, CDISC or hypernormal) and whether or not a standard file is used.

The add-ons (date detector, key item detector and low frequency detectors) are complementary tools used to ensure the best quality in the final result, the de-identified database for publication.

REFERENCES

[1]Guide to the De-identification of personal health Information, Khaled El Emam, CRC Press.
[2]Anonymizing Health Data, Khaled El Emam & Luk Arbukle,O'Reilly.
[3]Data De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach, TransCelerate Biopharma Inc. – http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/CDT-Data-Anonymization-Paper-FINAL.pdf
[4]PP22-KeyDAN: Risk Minimization in Data Anonymization – Phuse 2016 (Stéphane Chollet, Mathilde Laffitte).

ACKNOWLEDGMENTS

Cathy Scoupe, Head operations Belgium, Keyrus Biopharma
Jean Fernandez, Biometry expert, Keyrus Biopharma (JFE Consulting)
Mary Stimmel-Peeters, Clinical Research Associate, Keyrus Biopharma

CONTACT INFORMATION (HEADER 1)

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

Stéphane Chollet
Keyrus Biopharma
18/20 rue Clément Bayard
F-92300 Levallois-Perret
Email: stephane.chollet@keyrus.com
Web: www.keyrusbiopharma.com