PhUSE 2015

Paper SD06

De-ID: 'The What' and 'The How' of de-identifying your clinical trial data

Authors:
Gavin Winpenny, Business & Decision Life Sciences, Brussels, Belgium
Stephen Bamford, Business & Decision Life Sciences, Brussels, Belgium

Introduction

With regulatory agencies, researchers, journals and pressure groups all advocating greater transparency in clinical trials, the methodologies for successful anonymization of clinical data are an increasingly important consideration for organizations involved with study data.

This presentation will explore what is de-identified data, and what are the requirements you need to be aware of when anonymizing data? When planning to anonymise your data, what are the issues and pitfalls that await you and when you get to the end, how can you be sure you achieved what you set out to do?

The De-ID® application, which is already being used in large pharma, provides the tools to support a data professional through de-identifying their clinical data from single studies to drug projects, across CDISC (SDTM and ADaM) as well as organizations' legacy data standards.

The Challenge: Data Sharing - Transparency with Privacy

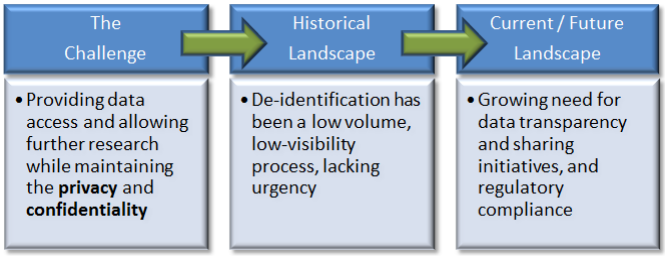

Figure 1 – The Challenge in Data Sharing

The pharma industry faces significant challenges when dealing with clinical trial data de-identification. Substantial investments within organizations to address the requirement to provide data access and allow further research while maintaining the **privacy** and **confidentiality of the patient.**

Looking at pharmaceutical companies, there are four main issues that they will need to face.

- Need to address variability in source data
- Reduce effort, time and cost
- Minimise risks to the privacy and confidentiality of research participants
- Ensure compliance with data privacy legal requirements

Some of these, such as the data privacy legal requirements, are very complex and are not yet fully defined. Furthermore, data privacy laws differ greatly between countries, so a company needs to ensure that releasing data in one country does not contradict with the laws in another where patients may have been based.

What is Anonymization/De-Identification?

Anonymization refers to severing a data set from the identity of the data contributor in a study to prevent any future re-identification. Clinical trial patient data falls under Protected Health Information (PHI) requirements, and the de-identification of patient data covers not just the more obvious identifiers of a patient, such as subject's race, ethnicity, age, date of birth, but also disease population, geographical location, etc., that singly or in combination could allow a subject to be re-identified.

When defining the Privacy Rules that should to be applied to satisfy de-identified data requirements for PHI, the U.S. Department of Health & Human Services (DOHHS) applies the following standard definition:

> §164.514(a) Standard: de-identification of protected health information. Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information.

To support the implementation of this principal, the DOHHS has developed two methodologies:

1. "Expert Determination" method
2. "Safe Harbor" method

The DOHHS has determined that satisfying either method would demonstrate that a covered entity has met Section 164.514(a) of the HIPAA Privacy Rule, with the de-identified health information created no longer protected by the Privacy Rule because it does not fall within the definition of PHI.

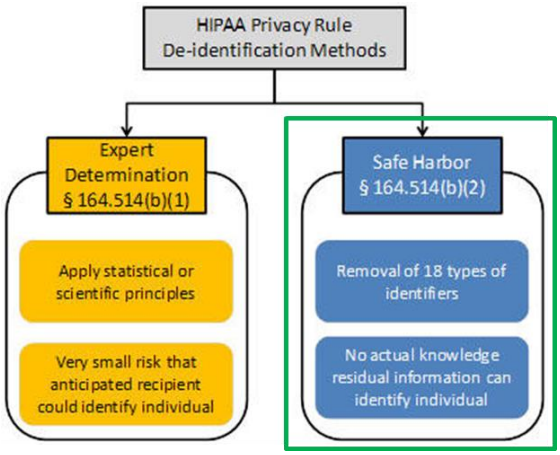

Figure 2 - HIPPA Privacy Rule: De-identification Methods

§164.514(b) Expert Determination method

Under the Expert Method a covered entity may determine that health information is not individually identifiable health information only if:

1. A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
   (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
   (ii) Documents the methods and results of the analysis that justify such determination

§164.514(c) Safe Harbor method

Under the Expert Method, the following 18 identifier types for the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names;
(B) All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:
   (1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and

(2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.

(C) All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
(D) Telephone numbers;
(E) Fax numbers;
(F) Electronic mail addresses;
(G) Social security numbers;
(H) Medical record numbers;
(I) Health plan beneficiary numbers;
(J) Account numbers;
(K) Certificate/license numbers;
(L) Vehicle identifiers and serial numbers, including license plate numbers;
(M) Device identifiers and serial numbers;
(N) Web Universal Resource Locators (URLs);
(O) Internet Protocol (IP) address numbers;
(P) Biometric identifiers, including finger and voice prints;
(Q) Full face photographic images and any comparable images; and
(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section (*Re-identification**) ; and

Additionally, the covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

**** Re-identification***

The implementation specifications further provide direction with respect to re-identification, specifically the assignment of a unique code to the set of de-identified health information to permit re-identification by the covered entity.

If a covered entity or business associate successfully undertook an effort to identify the subject of de-identified information it maintained, the health information now related to a specific individual would again be protected by the Privacy Rule, as it would meet the definition of PHI. Disclosure of a code or other means of record identification designed to enable coded or otherwise de-identified information to be re-identified is also considered a disclosure of PHI.

De-ID® and Safe Harbor methodology implementation

For purposes of this paper, the *Expert Determination* method is out-of-scope, and the remainder of this paper will concentrate on the *Safe Harbor* method, as this is the focus of the De-ID application.

Increasingly companies are now following the same approach with an emerging de facto industry standard:

In granting of access to data, there is some difference in approach across the industry; some companies release the data under a strict licencing agreement, while others will grant access to a closed system where the researcher can explore the data.

Figure 3 – De facto Industry Standard Approach

The issue remains of how to de-identify the data, such that patient confidentiality is maintained and the data still carries a high level of research value? The De-ID application applies the Safe Harbor rule-set to support organizations in enabling them to effectively and efficiently anonymize their data.

Commonality and Variability

When it comes to the approach pharmaceutical companies take to de-identify clinical data, there is a large amount of similarity between companies, but also variability in their approach to Safe Harbor implementation

Common Rules Across Sponsors:	***Varying Rules Across Sponsors:***
• Recode Patient ID and Site ID consistently • Delete or empty all free text (verbatim) fields • Don't retain ages > 89	• Dates (i) Apply Random offsets - used by about 60% of the companies that we are aware of; (ii) Replace with calculated Study Day - also commonly used, but does remove some data patterns (e.g. seasonality) • Dates of Birth (i) Replace with age (ii) Delete month and day • Re – identification index (key) Management The management of the key that is used to randomly adapt the patient's data can be different, with some organizations choosing to retain it, whilst others destroy it.

De-I.D. - Process Strategy Principals

In recognition of this variability, the BDLS De-ID solution accommodates variability between trials in terms of data structures and de-identification strategies.

Understand the Data

Most trials within an organization will share common data structures and domains, but each clinical study will have its unique aspects. The application of anonymization rules should **always be done with in context of the study**, and it is important to develop a **strong understanding and knowledge of the data, metadata and analysis** for a clinical study to enable **effective de-identification**.

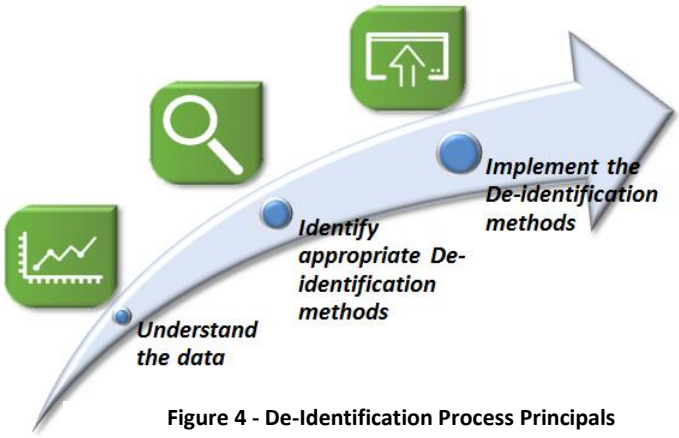

Figure 4 - De-Identification Process Principals

Identify appropriate De-identification Methods

Develop a detailed de-identification plan based on the metadata for each individual clinical study fully and document the de-identification functions to be applied to the applicable variables and records

Implement the De-identification Methods

A metadata-driven approach automates the application of the specified de-identified methods for **efficient de-identification**, with automatic generation the de-identification documentation including a description of the function used for each variable or record and the de-identification methods and parameters. Finally, the

destruction of all elements of the seed keys and commendation ensures complete anonymization of the patent data.

De-I.D. Software Approach

De-I.D. - Preparing anonymization instructions (rules file)

As an active member of this PhUSE De-Identification Working Group, BDLS participated along with 20+ pharma companies and CROs in the generation of the *De-Identification Standard for CDISC SDTM 3.2*. The goals of the Working Group were to facilitate the assessment of direct and quasi identifiers in CDISC datasets and to ensure consistency in de-identified data shared across sponsors.

Observation_Class	Domain_Prefix	Variable_Name	Variable_Label	Type	Direct_Quasi_Identifier (Direct/Quasi)	DI_Primary_Rule	DI_Alternative_Rules	DI_Comment
Interventions	CM	CMSPID	Sponsor-Defined Identifier	Char	Direct	Recode ID variable	Remove	In case, no direct identifier (e.g. CRF page number, lab sample number, etc.)
Interventions	CM	CMTRT	Reported Name of Drug, Med, or Therapy	Char	Direct	Remove	Review and only redact values with personal information	Other dictionnary coded variables being permissible in CM, If no dictionnary coded variables are available, the dataset will be useless without CMTRT and alternative rule should be used in this case. Note that marketing drug names may change from one country to another and should country be elevated to continent, CMTRT should also be checked if kept to verify that values do not reveal country.
Interventions	CM	CMMODIFY	Modified Reported Name	Char				
Interventions	CM	CMDECOD	Standardized Medication Name	Char				CM may include data that is linked with a sensitive diagnosis. It is the responsibility of the sponsor to decide on how to address such information.
Interventions	CM	CMCAT	Category for Medication	Char	Quasi Level 2	Keep		
Interventions	CM	CMSCAT	Subcategory for Medication	Char	Quasi Level 2	Keep		
Interventions	CM	CMPRESP	CM Pre-Specified	Char				
Interventions	CM	CMOCCUR	CM Occurrence	Char				
Interventions	CM	CMSTAT	Completion Status	Char				
Interventions	CM	CMREASND	Reason Medication Not Collected	Char	Direct	Review and only redact values with personal	Remove	Importance of the variable depends on the context. This information is not recoded anywhere else and can have an important data utility value for

+1300 variables

Figure 5 - PhUSE De-Identification Standard for SDTM 3.2 (Variable De-identification rule set)

The De-Identification Standard provides guidance on usage definitions, decisions, and rules, along with a highly detailed list of over 1300 SDTM variables, with Primary and Alternate Rules that could be applied for de-identification, as well as guidance notes on their application. The knowledge and experience gained in this participation has been incorporated in the De-ID application, and the tool is capable of feeding off the working group's De-Identification Standard.

Stage 1 - Preparing anonymization instructions

In addition to supporting SDTM 3.2, the application is highly flexible, with the capability to handle previous versions of SDTM, ADaM standards, and legacy data standards, via bespoke study specific rules files.

De-ID allows for up to 10 data source folders to be included in the de-identification process, with output datasets saved to specified target location(s).

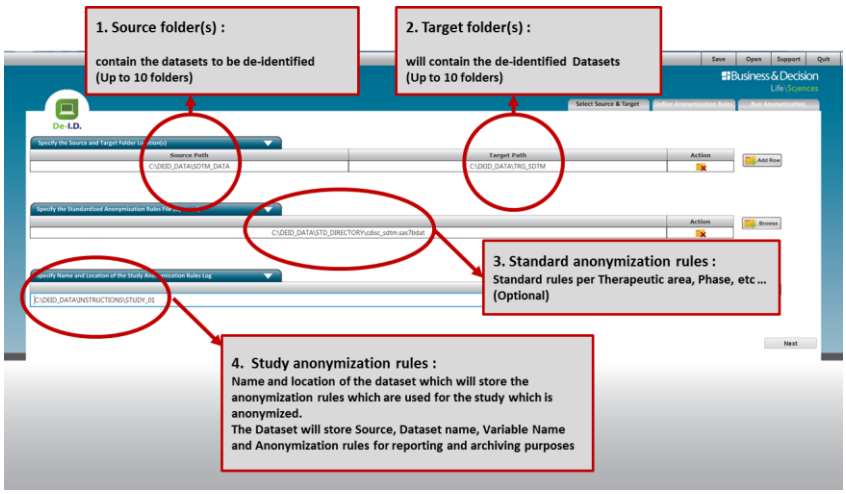

Figure 6 - De-ID's "Select Source and Target" screen

Standard and Study specific Anonymization rules files detail the de-identification function rules to be applied to a study's variables and observations.

Removing personally identifiable information (PII) from the dataset

The appropriate application of rules is important and should be applied with sensitivity to the data and the analysis. For example Country is an important Level 1 quasi Identifier, and in order to decrease residual risk (i.e. in the case of proactive data de-identification), by default, it is advised that Country be aggregated to continent as primary rule unless critical to the analysis (e.g. Country is a fixed-effect factor in a statistical model and the results cannot be reproduced).

The alternative rule is to keep Country as-is. In particular the rule described in HIPAA within the Safe Harbor method for geographical location (Rule B) is based on an empirical analysis performed by the US Census Bureau, and may not be applicable globally. Should this approach be adopted, then Site ID and Investigator ID need to be removed because a frequency analysis would likely reveal the most highly recruiting site in a country/region (which by definition would include many of the participants). The alternative rule is to recode Site ID and Investigator ID, if required for the analysis and in this case, it should be considered within the risk assessment.

Stage 1 – Defining Anonymization Rules to be Applied

To enable this, De-ID User Interface has seven De-ID Anonymization Rule Functions (more can be added) that can be applied to data at the **variable or record level** (through the use of 'Where…' clause Variable/ Values).

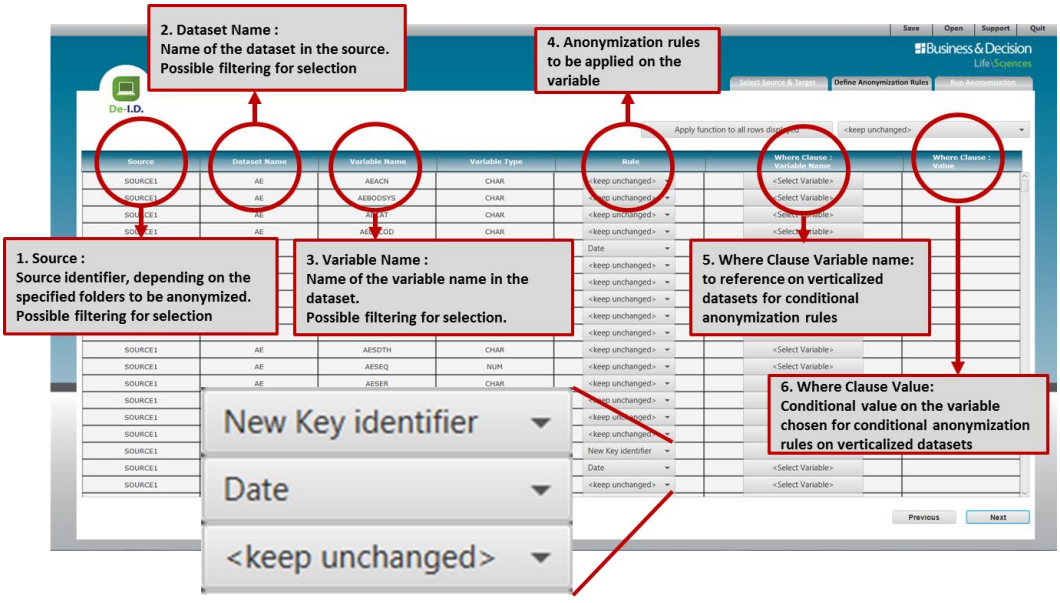

Figure 7 - De-ID's "Define Anonymization Rules" screen

De-ID Anonymization Rule Function:	***Description of De-ID Rule Function***
1. New Key Identifier variable (*SUBJ_ID* Function)	The function *SUBJ_ID* can be allocated to one variable only (subject identifier). The variable associated with the function *SUBJ_ID* will have all its values replaced by new values stored in the Master de-identification table. • If **KEEP_SORT** is set to YES, then the final

	table maintains its initial sorting. • If **KEEP_DEID** is set to YES, then the final table will contain both new and initial values. For a variable named xxx, the new values are in the variable xxx and the initial values in the variable xxx_pr.
2. Variable Anonymization (*VAR_ANOM* function)	The variable(s) associated with the function *VAR_ANOM* will have all values replaced by new values stored in the Master de-identification table. If **KEEP_DEID** is set to YES, then the final table will contain both new and initial values. For a variable named xxx, the new values are in the variable xxx and the initial values in the variable *xxx_pr*.
3. Date variables (other than date of birth) (*DATE* function)	The variable(s) associated with the function *DATE* will have the date value obfuscated by replacing the initial date with a newly computed date. The integer value **RAND_DAY** which is randomly computed between the bounds **LL** and **UL** is added to the initial date. **RAND_DAY** can be positive or negative depending on the bounds set by the user. The **RAND_DAY** value is stored in the table DEID_TABLE_*xxx**. A single **RAND_DAY** value is generated for each subject identifier. For a subject, all dates associated with function DATE are computed consistently, i.e. the **RAND_DAY** value is added to each date associated to DATE for each patient. • Acceptable date formats: date9., date in Character format, ISO8601 (yyyymmddThhmmss). The time will not be part of the anonymization process and will remain unchanged. • For dates reported in e.g. 3 fields (there may also be other ways how dates were stored over time), pre-processing will be required in order obtain a date in ISO8601 format in a single character field. • Partial dates: A missing month/day will be replaced by: ○ Middle day in month if month is known: 14 or 15 (Feb), 15 (Apr, Jun, Sep, Nov), 16 (Jan, Mar, May, Jul, Aug, Oct, Dec)

	○ Middle day in year if day and month missing, i.e. 30/06 (30th June)
4. Date of birth variables (*DOB_ANOM* function)	Variables associated with the function *DOB_ANOM* are obfuscated in exactly the same way as the variables associated with function. In addition, instead of displaying the entire date, with Days and Months, only the year of birth is displayed as final result. In the situation where the DE_ID function AGE is associated with at least one variable and the macro variable AGE_UNIT is set to YEAR by the end user, then the table DEID_TABLE_*xxx* will contain an additional variable DEID_AGE which contains the maximum value of age, per subject**.
5. Age variables (*AGE* function)	If the AGE function can be associated to one or more variables, then the associated to variable(s) then the max value per subject, of the values found in the variables associated with AGE are stored in the table DEID_TABLE_*xxx*. The variable created in the DEID_TABLE_*xxx* is DEID_AGE. In the final tables, if at least one subject has a value for age >89 and the macro variable AGE_UNIT is set to YEAR, then, for each table containing a variable associated with the function age: • A **new variable AGE_CAT** is created. This new variable will contain: ○ "> 89" for subjects older than 89 years ○ "=< 89" for the other subjects. ○ Empty for subjects having no age. ○ For subjects older than 89 years, the value for the variable associated with age is set to empty. For other subjects, the initial value is kept.
6. Variables to be dropped (*DROP* function)	The variable associated with the *DROP* function will be removed from the table. The target table will not contain the variable anymore.
7. Variables to be set to blank (*BLANK* function)	The variable associated with the function BLANK will be put to a missing (null) value in the target table, i.e. **blank** for character variables and **.** for numeric values.

* *xxx*: variable associated with subject identifier by using the function SUBJ_ID.

** Maximum value: in case more than one variable is associated to the AGE function: the program looks for the max value of the age, per subject

Stage 2 – Generating master anonymization tables

The study source data are combined and study-specific Anonymization instructions file applied to create Temporary Master Anonymization Tables.

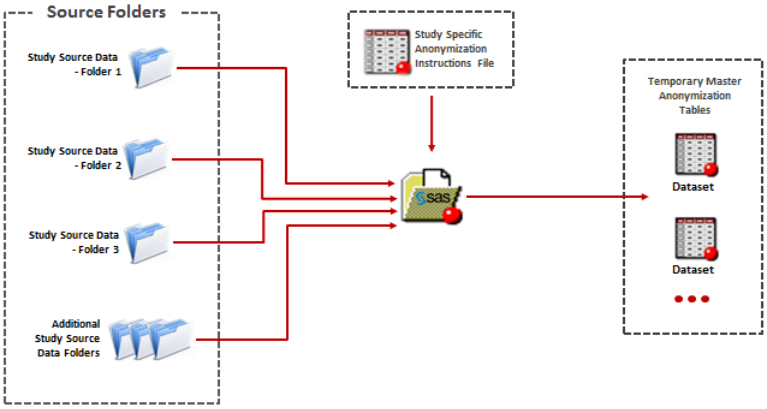

Figure 8 - Preparing Anonymization instructions

To aid review, the Temporary Master Anonymization Tables contain the original variables and values as well as their anonymized versions. This allows the user to check and confirm that the appropriate rule has been applied to a variable / record level prior to finalizing the anonymization of the data.

Stage 3 – Generating anonymization (Temporary and Final Anonymization) tables

The De-ID User Interface provides radio buttons and entry fields to control the characteristics of the anonymization to be performed.

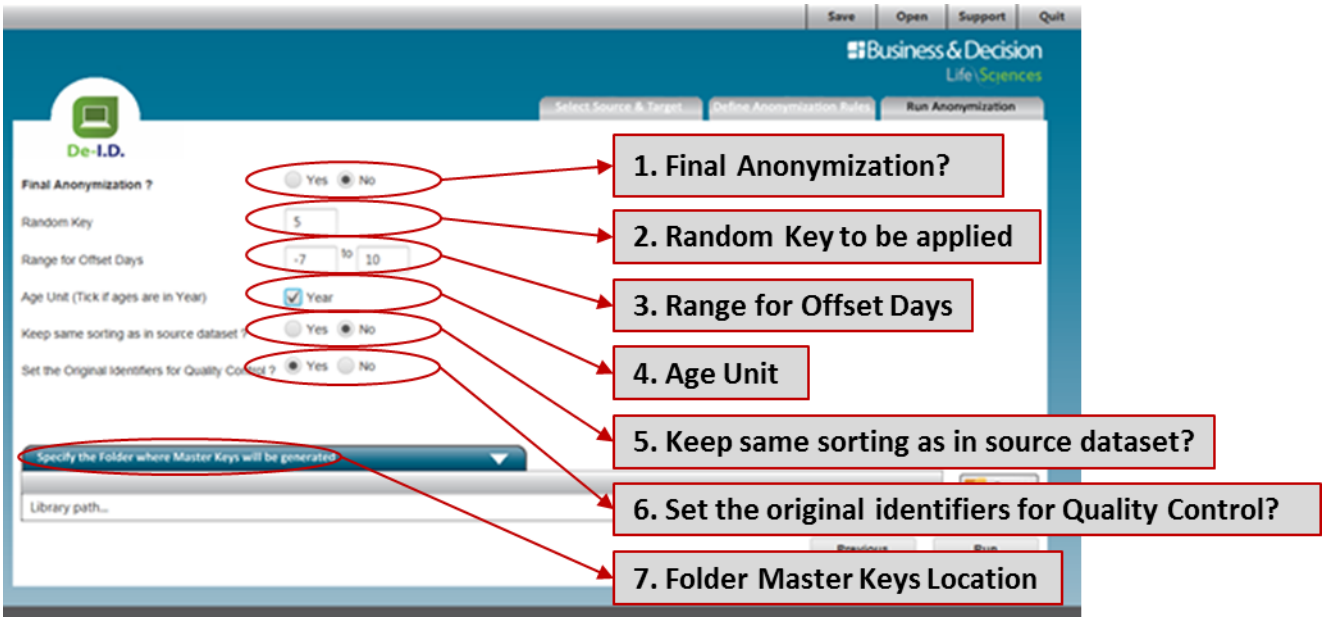

Figure 9 - De-ID's "Run Anonymization" screen

1. **Final Anonymization?**
   - **Yes:** all parameters (except Range for Offset Days) will automatically be set to final
   - **No:** used when, during anonymization preparation, the user wants to check and QC the anonymized tables (used before final anonymization)

2. **Random Key**

   Integer value used as seed for the master de-identification tables' generation.
   - If the value is zero (used for final anonymization), then the master keys can never more be generated.
   - When using a positive value, master keys can be generated with same values at the next run.

3. **Range for Offset Days (LL** and **UL**)
    - Integer values (lower and upper bound). Offset Days will be randomly assigned per subject within the specified range (value zero will never be selected).
    - Dates are anonymized by adding (or subtracting) the number of days selected to all dates to all dates associated with Date or Date of birth anonymization function, consistently for each subject across all datasets.

4. **Age Unit**
    - To be ticked if age variable(s) are expressed in Years

5. **Keep same sorting as in source dataset (KEEP_SORT) :**
    - **Yes :** final anonymized datasets keep same sorting as the source datasets
    - **No :** final anonymized datasets are sorted as per new identifiers ('No' is automatically selected for final anonymization)

6. **Set the original identifiers for Quality Control (KEEP_DEID):**
    - **Yes :** Initial values for anonymized datasets are kept in the final anonymized dataset, together with the anonymized ones , during QC process
    - **No :** Only final values are kept in the final anonymized datasets ('No is automatically selected for final anonymization)

7. **Folder Master Keys Location :**
    - Physical location where the master keys will be generated during anonymization. If the anonymization is Final, then the master keys will be automatically destroyed.

Conclusion

Effective de-identification of a study's data should always be considered in the context of the study data and the rules to be applied must be appropriate. The Benefits of De-ID's approach is it allows those involved in developing de-Identification processes to **efficiently assign robust, validated rule sets to their studies**.

De-ID's flexibility of design allows the automatic inclusion of SDTM 3.2 De-identification (selection of primary and Secondary rule should be done with knowledge of the study requirements), as well as legacy data standards.

The visualization of the data with the rule to be applied assists the user in confirming the rule is appropriate and should deliver the de-identification required. The ability to inspect the original data variable / value against the de-identified data value aids confirmation that the process has been followed correctly.

By following the steps of definition and documentation of de-identification requirements, care in their application through the DE-ID, user can avoid the risk of failing to correctly de-identify their clinical study data.

CONTACT INFORMATION

Contact the author at:
Stephen Bamford
Business & Decision Life Sciences
St Lambertusstraat 141 Rue Saint-Lambert
1200 Brussel – Bruxelles
Work Phone: +32 2 774 11 00
Fax: +32 2 774 11 99
Email: stephen.bamford@businessdecision.com
Web: www.businessdecision-lifesciences.com