Paper PP24

Data Anonymisation & Risk Assessment – Process Map and Automation Efforts

Lukasz Kniola, Biogen, Maidenhead, UK

ABSTRACT

In the expanding data sharing landscape, the expectation and demand to share clinical trial subject-level data is now undisputed. As techniques and standards are developed and collective experience is gained, trends and similarities emerge and offer opportunities to simplify, unify and automate the process of data anonymisation and risk assessment.

PhUSE Data Transparency WG is working on a deliverable which gathers, organizes, and proposes a process map, identifies areas which can be automated. This is achieved by breaking down the process into stages and further into individual tasks. This paper reports on the progress the team has made and builds on the work so far. It describes the steps necessary to transform source data to fully anonymised and documented package. It discusses individual tasks, assumptions and potential pitfalls. It explores opportunities for automation, proposes solutions and provides examples of how to automate individual tasks as well as the workflow.

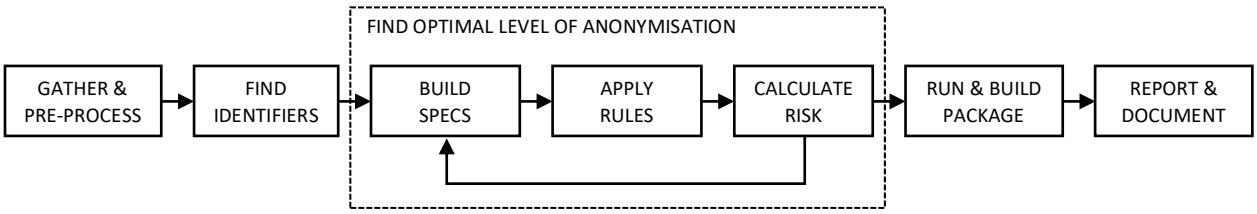
INTRODUCTION AND ASSUMPTIONS

This paper attempts to serve as a high level anonymisation recipe or a reference list for transforming datasets to anonymised data packages, while being general enough to be applicable in different scenarios.

Naming conventions, especially when discussing variables, are borrowed from the SDTM standard, but with some effort the process can be used for any format.

PROCESS MAP

The process of anonymising datasets – from sourcing data to preparing an anonimised package – has been split into several stages which, in turn, are broken down into individual tasks and points to consider. In its most simple form, the process map can be presented as below.



Each step is described separately with the goal of defining the inputs and outputs and considering tasks involved and opportunities for automation.

1. GATHER AND PRE-PROCESS DATA

DESCRIPTION

Before any anonymisation work can commence, source data needs to be identified. It is helpful to understand the relationships between datasets, their structure and formatting. Ideally, the source data would be in the SDTM format. However, there are steps that can be undertaken to make even legacy and unusual data more uniform. It can be argued that the time spent at this stage is likely to make the processing of the data much quicker and straightforward.

INPUTS

- Source data

OUTPUTS

- Clean datasets
- Dataset specs – containing [variable name, attributes]

TASKS

When working with SAS datasets the following transformations will be very useful in later stages:

3.1 Make sure that variables found in multiple datasets have the same attributes, especially variable type and format. When the same variable name is used for variables of different types (char, num) across datasets, this can cause serious problems.

3.2 Remove any SAS formats. Convert all formatted numeric variables to character (e.g. numeric SEX variable with codes: *1* and *2* should be converted to a character variable with actual values: *MALE*, *FEMALE*, etc.).

3.3 Convert all numeric dates to character variables in ISO8601 format.

3.4 As much as possible, replace non-standard values with controlled terminology values.

AUTOMATION

There is scope for automating a good portion of the above and any effort to make the data more uniform at this stage will be a great investment. However, although macros and functions can be programmed to fulfil specific tasks above, this step will likely require manual input and control.

2. FIND DIRECT AND QUASI IDENTIFIERS IN DATA

DESCRIPTION

When working with SDTM (or SDTM-like) data, the first step of finding direct and quasi-identifiers should be to check the variables against the PhUSE De-ID Standard for SDTM 3.2 spreadsheet (see references). At the end of the step, all variables should be classified as direct identifiers, quasi-identifiers or non-identifiers.

INPUTS

- Clean datasets (from step 1)
- Dataset specs – containing [variable name, attributes] (from step 1)
- PhUSE De-ID Standard for SDTM 3.2 spreadsheet

OUTPUTS

- Updated dataset specs – containing [variable name, attributes, **+classification**]

TASKS I – SDTM DATA

2.1. Merge the dataset specs, containing the list of datasets/variables with the PhUSE De-ID Standard for SDTM 3.2 spreadsheet. There are two levels of merging – by full variable name (e.g. RFSTDTC) and by partial name (e.g. --TERM).
This step will help find the vast majority of direct and quasi identifiers as well as recommended anonymisation rules to apply to the discovered identifiers.

TASKS II – USING VARIABLE NAMES

Additionally, identifiers can be discovered by considering the variable names and labels:

2.2. Find demographic variables: age, sex, country, region, race, ethnicity, etc.

2.3. Find body measurement values: height, weight, BMI, etc.

2.4. AE, ConMed, MedHist terms.

2.5. Date and time variables.

TASKS III – USING VARIABLE CONTENTS

Finally, identifiers can be found by checking the contents (values) of variables:

2.6. All the above.

2.7. Birth / death dates.

2.8. Free entry (comments) variables.

2.9. Variables potentially containing sensitive information (pregnancy, abortions, mental health, substance abuse, family, etc.).

In addition, consider information which may help infer values of quasi-identifiers.

2.10. Check laboratory analytes which can reveal sex of subject, pregnancy questions, unreported race may suggest subject being in a country where collecting this detail is not permitted, etc.

AUTOMATION

Task I can be automated with little effort. All ingredients are readily available.
Automating task II would require creating a comprehensive list of variable names which could be used in a similar fashion to task I. It will always require manual check, to make sure that unexpected variable names are caught (and added to the list).
Task III is certainly the most complex and is likely to require human review.

3. BUILD SPECIFICATIONS

DESCRIPTION

Once all direct and quasi-identifiers have been found, the next step is to prepare dataset specifications by assigning anonymisation rules to each variable across all datasets. Based on the classification of all identifiers and non-identifiers each variable should be assigned a corresponding rule.

INPUTS

- Dataset specs – containing [variable name, attributes, classification] (from step 2)
- PhUSE De-ID Standard for SDTM 3.2 spreadsheet

OUTPUTS

- Updated dataset specs – containing [variable name, attributes, classification, **+anonymisation rule**]

TASKS

Based on the type of identifiers, assign one of the following rules:

3.1 "DID" – recode ID – this can be achieved by creating a list of *original:anonymised* pairs to "translate" the ID values, or by hashing the IDs. Recoding of IDs must be applied consistently across all datasets, to allow for linking of records.

3.2 "KEEP" – retain variable as is, without any manipulation. NOTE: This rule will be applied to non-identifiers.

3.3 "DROP" – remove variable from dataset.

3.4 "CLEAR" – keep the variable but leave empty.

3.5 "OFFSET" – apply date offsetting technique (see PhUSE De-identification standards for CDISC SDTM 3.2, Date Offsetting Appendix).

3.6 "NEW" – add variable, additional definition is required.

3.7 "DERIVE" – apply transformation to the value of the variable:

- 3.7.1 "AGE_BANDS" – convert age value to an age band. Examples may include 5-year bands (e.g. *31-35*), 10-year bands (*30-39*), splits *(<55 / >=55*), etc.
- 3.7.2 "AGE_IN_BAND" – similar to 3.7.1 except the resulting value is a random number from within the band (e.g. *33 -> [31-39]* -> random_from_band: *37*)
- 3.7.3 "AGE89" – age values >89 are recoded as *89+*
- 3.7.4 "COUNTRY_POOL" – elevate country information to regions, continents, etc.
- 3.7.5 "LOW_FREQ_POOL" – change values with low frequencies to *OTHER*. E.g. RACE: *WHITE / OTHER*.
- 3.7.6 "REDACT"- replace part of full value with **redacted**.*
- 3.7.7 "OTHER" – non-standard formula for dealing with one-off cases.

All the rules can also be applied to a subset of records. A conditional rule should be applied to values of --ORRES, --STRESC, --STRESC, AVALC, AVAL, etc., based on the values of --TESTCD or PARAMCD (for SDTM/ADaM datasets, or equivalent for non-CDISC datasets).

AUTOMATION

Specifications can be generated automatically based on the classification of variables from step 2.

4. APPLY ANONYMISATION RULES

DESCRIPTION

Once all direct and quasi-identifiers have been found, the next step is to prepare dataset specifications by assigning anonymisation rules to each variable across all datasets. Based on the classification of all identifiers and non-identifiers each variable should be assigned a corresponding rule.

INPUTS

- Clean datasets (from step 1)
- Dataset specs – containing [variable name, attributes, classification, anonymisation rule] (from step 3)

OUTPUTS

- Anonymised datasets

TASKS

4.1. Apply individual rules to each variable to transform the values accordingly.

4.2. Output anonymised datasets, ensuring all values which require transformation have been replaced or removed.

AUTOMATION

Since the specifications build in steps 1, 2, and 3 contain all relevant details, i.e. variable attributes and anonymisation rules to be applied, it's possible to build a macro or function which iterates over the variable list and applies rules variable by variable.

A "catch all" rule should be added to make sure that any variable in the dataset without a corresponding rule assigned triggers a warning and such variable is dropped by default.

5. CALCULATE RESIDUAL RISK

Risk of re-identification is calculated to confirm if the applied transformations are sufficient for the data to be shared and reused safely, and participants' privacy is respected and secure. This is achieved by comparing the calculated residual risk of re-identification to a pre-specified threshold.

For detailed description of risk quantification process refer to (PhUSE 2016, DH09 - Calculating the Risk of Re-Identification of Patient-Level Data Using Quantitative Approach, Kniola, 2016).

INPUTS

- Anonymised datasets (from step 4)
- Reference dataset – for journalist risk calculation (optional)

OUTPUTS

- Risk calculation report

TASKS

At high level, the following tasks are required for full risk assessment:

5.1. Select quasi-identifiers for risk calculation – not all quasi-identifiers are necessarily taken into account. Dates can be offset or truncated (to year or year-month) without having to be included in the risk assessment.

5.2. Assess risk of attack:

5.2.1. Risk of deliberate attack,

5.2.2. Probability of inadvertent re-identification,

5.2.3. Risk of data breach,

5.3. Calculate record-level individual values of risk of successful re-identification in the event of an attack. Find average and maximum risk across the dataset.

5.4. Calculate overall risk of successful re-identification using results from 5.2 and 5.3.

5.5. Assess k-anonymity and uniqueness in the dataset.

AUTOMATION

Although the list of quasi-identifiers will be different from one project to another, the algorithm of finding individual risks (and consequently dataset-level risk) given the list of quasi-identifiers will always be the same. Once a macro or function is programmed with parameters built into the design, it can be re-used with little or no modification.

R. SELECT OPTIMAL LEVEL OF ANONYMISATION

The goal of anonymisation is to find the "Goldilocks" level of data transformation which will ensure that the risk of re-identification is sufficiently low (below a pre-defined threshold), but not lower than necessary. Being too conservative will inevitably result in unwarranted data utility loss. In practical terms, this means finding a set of rules which results in the residual risk being just below the threshold.

Moving along the process map, we have so far classified the variables (step 2), assigned a set of rules (step 3), transformed the datasets according to those rules (step 4), and calculated the residual risk of re-identification (step 5). However, there is no such thing as "one size fits all" in the world of data anonymisation, and optimal set of rules depends on the input data – number of records, number and type of identifiers and their values.

Each quasi-identifier can be assigned different rules with different levels of detail retained. Consider the following examples – AGE, COUNTRY, SEX – and rules which can be applied, how they affect residual risk and retention of detail (data utility).

Example 1: **AGE** – original value: **32**

Rule	Value after rule applied	Level of detail retained	Residual risk
KEEP	32	highest	highest
AGE_BANDS (5-year bands)	31-35		
AGE_BANDS (10-year bands)	31-40	↓	↓
DROP	-	lowest	lowest

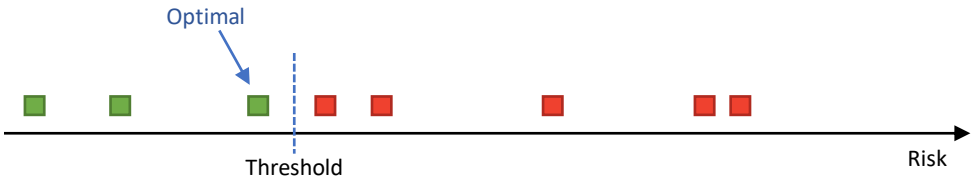
Example 2: **COUNTRY** – original value: **Poland**

Rule	Value after rule applied	Level of detail retained	Residual risk
KEEP	Poland	highest	highest
COUNTRY_POOL (regions)	Central Europe	↓	↓
DROP	-	lowest	lowest

Example 3: **SEX** – original value: **Female**

Rule	Value after rule applied	Level of detail retained	Residual risk
KEEP	Female	highest	highest
DROP	-	lowest	lowest

Each identifier will have at least the options of being kept or dropped. Some can have multiple levels of detail retention. Those levels are arbitrary and the choice and number depends on the preference and requirements of the requestor and/or the analyst.
To make sure that the optimal set of rules is found, it is advisable to calculate the risk for all practical scenarios and choose the one that is below – and closest to – the threshold, as illustrated below.



TASKS

R.1. Create list of combinations of anonymisation rules.

R.2. Run steps 5, 6, 7, iterating over the list.

R.3. Select optimal scenario (with residual risk just below the threshold).

AUTOMATION

The number of combinations will depend on the number of quasi-identifiers and number of possible rules that can be applied to them.

E.g. In a dataset with three quasi-identifiers:

- AGE – 4 levels,
- COUNTRY – 3 levels,
- SEX – 2 levels,

the number of combinations will be: 4*3*2 = 24. This may seem high, especially that a real-life dataset is likely to have more quasi-identifiers. However, utilizing the spec generation macro from step 3, rule application macro from step 4, and risk calculation macro from step 5, it is easy to iterate over all combinations (sets of rules), calling the macros to rerun the steps. The result should be a list of all possible combinations with a risk value calculated for each of them. The ultimate goal is to find the scenario where risk is just below the threshold.

6. RUN PRODUCTION & BUILD ANONYMISED DATA PACKAGE

Once the optimal set of rules has been found, the next step is to build the package.

INPUTS

- Clean datasets (from step 1)
- Final dataset specs – containing [variable name, attributes, classification, anonymisation rule] – with the optimal set of rules

OUTPUTS

- Final anonymised datasets in the desired format
- Final dataset specs
- Anonymisation report
- All elements of the deliverable compressed in a single zip file

TASKS

6.1. Using the optimal set of anonymisation rules (from step R) rerun steps 3 and 4, saving the final specification.

6.2. Rerun step 5, saving the final risk calculation results for documentation purposes.

6.3. Outputs of 6.1 and 6.2 form the basis for the anonymisation report which describes assumptions, steps taken to sufficiently reduce the risk of re-identification and confirmation of the residual risk values being acceptable.

6.4. Convert the datasets to the desired format – sas7bdat, xpt, csv, Dataset.xml, JSON, etc.

6.5. Compress all into a single zip file and password protect. It's suggested to follow a naming convention for the zip file which includes the project name and creation date (e.g. rq001-2019-11-10.zip).

6.6. Securely store zip password for future reference.

AUTOMATION

Assuming relevant macros have been previously programmed for steps 3, 4, and 5, creating remaining macros or functions to convert the datasets and zip the package should, at this stage, be trivial.

7. REPORT AND DOCUMENT

Apart from the anonymised datasets and metadata describing the structure of the data and the classification of the variables, the third equally important output is the anonymisation report, which should include:

- List of identifiers (direct and quasi)
- Classification of all variables and final rules applied to them
- Description of risk calculation methods (choice of prosecutor/journalist risk, selection of reference population, etc.)
- Values of risk of attempt (deliberate, inadvertent, breach, overall)
- Results of risk calculations – to avoid listing all scenarios and respective risk values, it is advisable to report the risk value of the original dataset (prior to any rules being applied) and the final dataset (with final set of rules applied).
- Description of impact on data utility

Individual steps can contribute to the anonymisation report and outputting parts of report can be programmed into macros and functions for the process steps. They can be ultimately collated programmatically to form the bulk of the report, with limited human input required to fill in any gaps.

CONCLUSIONS AND NEXT STEPS

Anonymising datasets is a complex task and involves a thorough understanding of the process. Having a comprehensive reference may serve as a recipe which lists individual tasks to consider when anonymising data, helps to avoid missing crucial steps (by design or omission) and lists the desired inputs and outputs at different stages of the process.

PhUSE De-Identification Working Group is – at the time of submitting this paper – working on a process map which provides as much detail as possible on individual tasks, considerations, pitfalls as well as helps ascertain the potential for automation on a task, step and even cross-step level. The ambition is to create a comprehensive document, reviewed by the community and accepted as the universal base for the dataset anonymisation process. The deliverable is planned for autumn/winter 2019.

REFERENCES

Kniola, L (2016). *Calculating the Risk of Re-Identification of Patient-Level Data Using Quantitative Approach.* PhUSE Annual Conference, 2016.

Institute of Medicine. (2015). *Sharing Clinical Trial Data Maximizing Benefits, Minimizing Risk.* Washington, DC: The National Academies Press.

Pharmaceutical Users Software Exchange (PhUSE). De-identification standards for CDISC SDTM 3.2.

CONTACT INFORMATION

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

Lukasz Kniola
Biogen Idec Ltd
Innovation House
70 Norden Road
Maidenhead, Berkshire SL6 4AY
UK
Email: lukasz.kniola@biogen.com