PP11

A process comparison between:

ARX Data anonymization and risk assessment

PHUSE Data anonymization and risk assessment automation

Authors: Valentina Aguilar, Michael Healey, Mario Salas

ICON Biostats Consulting Services, Mario.Salas@iconplc.com

In the last decade, the unregulated disclosure of personal data has prompted European Medicines Agency, Health Canada, and other competent authorities to issue guidelines on sharing and publishing clinical trial data. In 2018, the FDA, as part of the broader Clinical Trial Transparency effort, implemented a voluntary program to evaluate disclosure of information provided by clinical study reports. The pilot emphasizes the need for anonymization and disclosure standards and acknowledges the ongoing independent effort by PHUSE As an example, CDER's Office of New Drugs will consult with stakeholders on portions of data to be posted, and new guidelines will be established. Numerous data-sharing entities have been established with the vision of accelerating drug discovery and development. Examples are C-Path, a nonprofit public-private partnership with the FDA, which aims to reduce costs by data sharing; TCIA Cancer Imaging Archive, that de-identifies medical images for public download; CSDR consortium, that facilitates the access to patient -level data; and NCI controlled-access database, that stores clinical trial data generated by the NCI.

How can anonymization and risk assessment be accomplished?

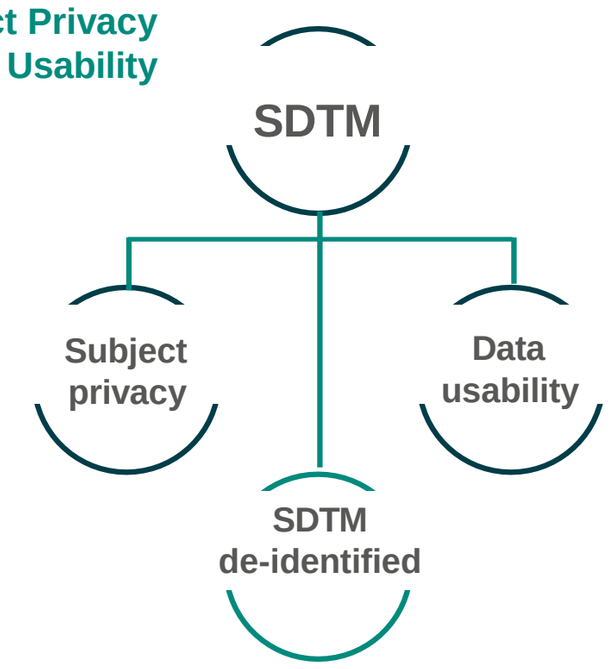
Analyse the context	Select variables for quantitative risk assessment	Apply PHUSE guidance to create transformation specifications
Communi-cate recommend-dations	Adjust specs per data- sharing agreements or regulatory agency	Use SAS progs or anon tool, transform data per risk assessment
De-identify SDTM	Validate de-identified SDTM, Produce Risk report	Secure Mapping transformati on Destroy original data

Why is anonymization of SDTM needed?

Anonymization is a balance between subject Privacy and Data Usability

Current EMA and HC regulatory requirements call for document and sometimes data anonymization.

Public and controlled/limited access of clinical trials data sharing site require data anonymization while retaining data usability



What automated methods exist to anonymize data?

There are several companies that provide data anonymization and document anonymization software.
A major advantage is the capability of harmonisation between mappings used in anonymizing datasets and those mappings used in de-identification of documents.

Proprietary software

ARX Data Anonymization Tool: supports several models, good graphical interface, and can be downloaded

Amnesia: Has an online version, supports 2 main models

U-ARGUS: based on R, range of methods

Open source

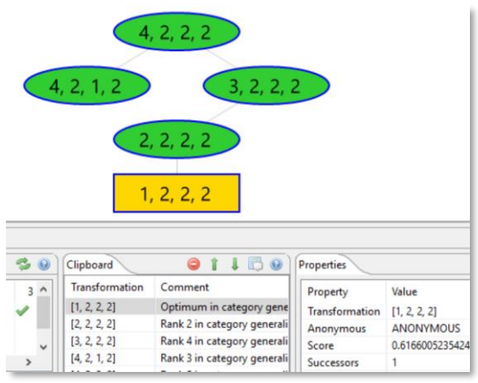
PHUSE Data anonymization and risk assessment automation (PHUSE Data Transparency Group 2019) provides clear guide to risk assessment with programming code examples, which illustrate how to implement k-anonymization models and risk assessment in SAS, Julia, R environments.

In house programs

What we did

In this comparison, with the purpose of sharing data between a sponsor and a data-sharing consortium and using k-anonymity model (value k>=5) to reduce risk of de-identification, we anonymized SDTM datasets using two tools:
1. ARX Data anonymization tool which produced anonymized datasets and a risk assessment report
2. In-house programs created by following PHUSE sample programs and appropriate mappings

ARX[2]



Pros

- Multiple models available: k-map, l-diversity, t-closeness
- Cost/benefit analysis considers marketer attacks
- Visual interface which displays solution space, risk, useful stats, and much more
- A library of hierarchy spreadsheets can be built and shared between studies

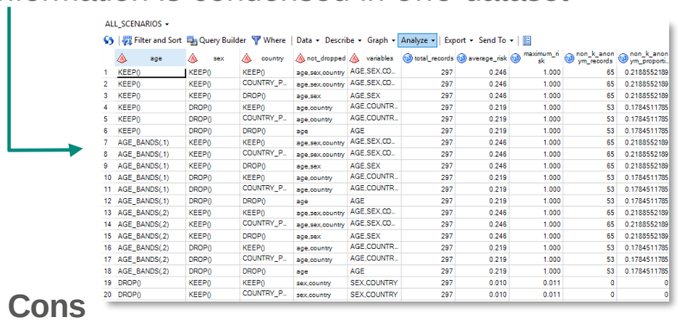
Cons

- Additional system setup which can take much overhead
- Requires pre and post processing in SAS
- Mapping must be defined on spreadsheets with specific formats
- Spreadsheets must be kept as part of production environment
- Results can appear out of a black box and are difficult to interpret

In House PHUSE[1] Programs

Pros

- Uses existing single SAS prog. environment.
- Anonymization scenarios are clearly defined in main code and outputs have easy-to-understand information
- Solutions can be selected based on risk/data usability
- Mappings are part of program and easily adjusted
- Other models like l-diversity can be added
- Information is condensed in one dataset



Cons

- No tool coordinated for anonymization documents
- No complex models in original sample code, programs and outputs must be validated

Example: SENSITIVE ATTRIBUTE RACE 2-DIVERSITY *For full SAS program follow QR code*

Given deidentified DM with classes indexed by variable i, find those with <3 different values of RACE and mask;

```
proc sql;
create table nondiver as select i,
count(distinct RACE) as cnt
from dm_deid group by i having
0<calculated cnt<3;
quit;

data dm_deidc;
merge dm_deid nondiver (in=a); by i;
if a then RACE="*****";
run;
```

Conclusions

For flexibility, systems set-up and setting of parameters and mappings, the PHUSE recommended SAS code was better than ARX. In addition to demographic pseudo-identifiers, non-pseudo-identifying free text fields that can identify dates and other personal information, like rare conditions, had to be masked using SAS in post risk assessment, post de-identification. The ARX tool handles k-map and other complicated risk-assessment models well. Other models such as t-closeness can be added to help evaluate risk from non-quasi-identifier fields containing sensitive information. The tool also provides a dynamic way of examining the anonymization solution space by providing graphs which compare the before and after frequency distributions of each attribute. Using the PHUSE recommended programs in the validation process provided understanding of the results in familiar terms of SAS programming. As a bonus, having intermediate data sets provides clarity that can be communicated with sponsors and regulatory authorities.

References

1. PHUSE Data anonymization and risk assessment, Version 1.01, access at https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Data+anonymization+and+Risk+Assessment+Automation.pdf
2. ARX – Data anonymization tool, access at https://arx.deidentifier.org/overview/ privacy-criteria/
3. PHUSE De-Identification Working Group: Providing de-identification standards for cdisc data models, 2015, access at https://www.pharmasug.org/ proceedings/2015/DS/PharmaSUG-2015-DS10.pdf

For more information, read the complete whitepaper.

ICONplc.com