

EUCAIM Wizard Tool (FORTH)

Summary: The EUCAIM Wizard Tool performs an analysis of data re-identification risks of imaging and clinical data that follow the EUCAIM CDM. It includes and uses an EUCAIM specific configuration of the ARX Data Anonymization Tool (biotools:arx), by supporting a wide variety of privacy and risk models as well as methods for analyzing the usefulness of output data.

Status:

Contacts: nikiforakik@gmail.com

Table of Contents

I. Purpose	2
II. Tool description for its conceptual validation	2
1. Name:	2
2. Contributor:	2
3. Area:	2
4. Tool description:	3
5. Data:	3
6. Methodology/performance:	3
7. Use: brief description of the tool's functioning (if it applies).	3
8. Input/output formats: description of the input/output of the tool at the validation stage (will help if some adaptation is needed).	3
9. Quantitative results: performance obtained during training of the tool (if applies)	4
10. Qualitative results: Provide some visual results (if available) of applying such tools.	4
11. Additional information: successful use cases, external resources (open code, papers...) licence, certification ... (if they apply).	5
III. Technical specifications	7
1. Data: In depth description of the data used to train the tool.	7
2. Methods: in depth description of the methodology used for its development including all data preprocessing.	7
3. Specific Technical information:	8
4. Traceability and monitoring mechanism.	8
5. Unitary tests: description of the tests implemented to verify the correct	

functioning of the tool.	8
6. Access restriction : do you have any access restriction to the source code or to the binaries of the tool?	8
7. Additional information for tool integration.	8
IV. Integration Validation	9
1. Communication channel for the helpdesk, technical support channel:	9
2. Most common errors:	9
3. FAQs:	9
4. User Manual:	9

I. Purpose

Purpose: The EUCAIM Wizard tool is based on ARX software which is open-source data de-identification / anonymization software providing scalability, usability and support for various anonymization techniques. It has been designed for anonymizing sensitive personal data, particularly structured (tabular) microdata like datasets containing demographic or biomedical information.

The main goal of re-identification risk analysis is to find the optimal balance between data security and utility. For this reason, information regarding a specific patient cohort organized in a tabular form is imported to initiate an analysis. The user sets a number of criteria and constraints to the permitted changes to be applied and the output is the modified list of metadata. The main goal is to propose actions to minimize risk of re-identification while preserving data utility to the largest possible extent. The actions are based on freely available software -ARX Ref- which has been configured for the specific requirements of the EUCAIM platform. The configuration takes into account the nature of provided data which can be imaging and non-imaging . The former contains data complying with the DICOM standards even after anonymization actions and thus hold at least the minimum necessary DICOM tags to maintain DICOM format integrity. The non-imaging data comprise clinically relevant information provided in a tabular form and regard information that complement imaging data to draft the patient's clinical profile in a specific time point or throughout the disease continuum. The combined information from imaging and non-imaging data is the main input for the ARX platform. The output is a structured report with initial risk estimation and a list of proposed actions for balancing security and usability for the specific patient cohort.

ARX is an open-source software for anonymizing sensitive data. It supports a wide range of privacy models (e.g., k-anonymity, l-diversity, t-closeness, differential privacy) and provides tools to evaluate the trade-off between privacy and utility. This document provides concise guidance for using ARX to minimize re-identification risks within large databases while preserving analytic value.

The tool offers the ability to apply changes that increase security of the de-identified cohort but also provided measurable evidence on the effect of the applied modifications to data

usability. In this way, overaggressive de-identification is prevented while achieving evidence based usability metrics, also outlining the data value and identity. The user is free to apply a number of proposed scenarios iteratively, until the optimal balance of security and usability is achieved, depending on the cohort size and the use case.

II. Tool description for its conceptual validation

The conceptual validation aims to ensure that a tool is aligned in its purpose and functionality with the pre-processing tools specified in EUCAIM T5.3. For this, the following documentation will detail **the necessity of the tool and its functioning**.

1. Name:

EUCAIM Wizard Tool

2. Contributor:

FORTH

3. Area:

Anonymization- Risk and Utility estimation

4. Tool description:

The proposed tool follows a multi-step process by identifying the degree of identification potential of each attribute provided by clinical data or imaging meta data. The degree of generalization and suppression is user-defined, as well as the anonymity requirements per model selected. Then the final step is to explore the solution space and iteratively define the optimum solution and register the suggestions to be applied to the dataset. The final stage is to modify the data by using the suggested actions from the previous steps.

By applying ARX based WIZARD's privacy models, carefully designing generalization hierarchies, and balancing risk with utility, organizations can significantly reduce re-identification risks while maintaining the analytical value of large datasets.

5. Data:

Any kind of tabular dataset, where the number of rows correspond to patient records and the number of columns corresponds to features which can be related to different categories, like demographic, laboratory examinations, medical conditions, therapies, etc.

6. Methodology/performance:

The tool was developed following a waterfall architecture. The main function of the data curator executes the following phases:

- (i) **Data / Meta data preparation & Import** which involves the preparation of data and identification of quasi-identifiers (QIs) of cohort, as well as the permitted levels of generalization (hierarchies), which can be made in the tool itself or can be imported as csv or excel files
- (ii) **the parametrization** of the solution space, i.e. identification of the anonymization model, attribute weighting, generalization/suppression limits
- (iii) **the exploration of the solution space and usability analysis** which is the stage where the graphical interface presents a number of different solutions within the constraints defined by the user, as well as contingency plans in order to finally conclude to the suggestions that best support the needs of the anonymization scheme of the specific patient cohort. The result is exported as a report documenting the modifications and the final step (work in progress) will be to integrate the result to the provided data set.

7. Use: brief description of the tool's functioning (if it applies).

Preparing the Data

The first action for data preparation is to create a tabular list of attributes for each patient in a tabular form where each tuple contains all the information of a diagnostic event, with imaging and/or non-imaging information. The whole patient cohort is a list of such rows with the same number of columns either populated or not. When the diagnostic event is an imaging process, the DICOM header is extracted and all the tags present after anonymization are combined with the available clinical information and structured in a patient -specific format and inserted as a separate row in an aggregative table containing the whole patient cohort.

After the complete cohort is combined into a tabular form, the user opens the ARX interface to import the csv or excel file. The user has the ability to inspect the set of attributes to be imported and clear any irrelevant columns at this initial stage. When accepted, the populated list of attributes is presented to the user and the user is ready to distinguish four different types of attributes with respect to the degree that they could violate security or compromise utility of the dataset. Thus, there are four main categories of data entries, classified with respect to their ability to disclose patients identity:

- Identify quasi-identifiers: Values that can be linked to external information and thus provide a link to the patient's identity.
- Distinguish identifiers: Values directly linked to patients identity and are prohibited to be retained.

- Sensitive attributes: Values yielding sensitive information regarding the patient or group of patients requiring strong protection, i.e related to socially vulnerable groups.
- Insensitive attributes: Attributes that can be ignored in the frame of a security-utility analysis, possibly not related to each individual patient's identity. They can be retained without change.

Summarizing the above:

- **Identifying:** must be removed or pseudonymized.
- **Quasi-identifying:** require transformation to prevent linkage.
- **Sensitive:** require strong protection (diseases, income, etc.).
- **Insensitive:** can be retained without change.

Indicatively, *Name* as Identifying, *ZIP Code* as Quasi-identifying, *Religion* as Sensitive, and *Echo Time* as Insensitive.

3. Defining Privacy Models

The software offers the ability to choose among different privacy models, depending on the specific needs of the cohort. The user has the ability to perform cohort risk analysis based on the following models:

- k-Anonymity: Ensures each record is indistinguishable from at least $k-1$ others based on quasi-identifiers.
- l-Diversity: Extends k-anonymity by requiring diversity in sensitive attributes within each group.
- t-Closeness: Limits how much the distribution of sensitive attributes within a group can diverge from the overall dataset.
- Risk-Based Models: ARX also supports prosecutor, journalist, and marketer risk scenarios.

For cohorts containing non-sensitive data, k-anonymity is recommended as it can structure an efficient and simple risk minimization schema. Open the *Privacy Model* panel → Add *k-Anonymity* → set $k = 10$. This means each combination of quasi-identifiers must appear in at least 10 records. For large databases, start with $k \geq 5-10$, then evaluate l-diversity or t-closeness depending on the sensitivity of the data.

4. Transformation Techniques

In order to configure the risk minimization strategy, a number of different techniques can be utilized as single or combined methodologies:

- Generalization: Replace values with broader categories (e.g., exact age → age ranges).
- Suppression: Remove high-risk values that cannot be anonymized effectively.
- Hierarchy Design: Define hierarchies for generalization (e.g., ZIP 12345 → 1234* → 123**).

The above mentioned methodologies are ingested by the tool in the form of “hierarchies” which are imported in an excel format or directly defined inside the tool in the specified editor space. Balanced hierarchies preserve utility while ensuring privacy. In the *Hierarchy Editor*, create levels: *Exact Age* → *Age Group* (e.g., 20–29) → *Decade* (e.g., 20s) → *Broad Category* (e.g., Adult). The user is able to define the extent of generalization or suppression actions that are considered plausible for each certain case.

5. Evaluating Risk

The WIZARD tool performs a risk analysis module to identify records at highest risk, examine possible actions to minimize risk while measuring usability metrics ensuring non aggressive anonymization that degrades data quality for subsequent use. The user is able to define the balance between security and usability by running the analysis and then navigating through the proposed scenarios which have been ranked according to their performance in security or usability tests. The user is also able to see statistical metrics of the cohort once a certain solution is selected within the proposed solution space, and to juxtapose it to the initial dataset. Adjusting privacy parameters and inspecting the results is an iterative process which can be continued until acceptable thresholds are met. For example, the *Risk Analysis* tab shows risk distributions (e.g., highest risk = 9%, average risk = 0.8%).

6. Preserving Utility

The utility aspect of the dataset after recommended actions for security improvement are quantified in a set of ARX’s utility metrics e.g., information loss, average equivalence class size). The user is invited to compare candidate anonymization schemes and is supported by a number of data utility visualizations to ensure statistical analyses remain valid. For example, the *Data Utility* view provides a graph comparing different anonymization results, allowing selection of the transformation with the best balance between risk and utility. The ultimate aim is to configure dataset attributes achieving the lowest re-identification risk that still allows your intended analyses.

Feature (Utility Metric)	Category	Explanation	Example Use Case
Discernibility	Record-Oriented	Penalizes records that belong to small equivalence classes and heavily penalizes suppressed records. Lower score is better (less information loss).	Used when maximizing the number of usable (non-suppressed) records is the main goal.
Non-Uniform Entropy	Attribute-Oriented	Quantifies the loss of mutual information between the original and generalized attributes. A complex measure of how much the original data distribution is distorted.	Useful for datasets where preserving the statistical relationships between attributes is critical for analysis.
Precision (Generalization-based)	Cell-Oriented	Measures the level of generalization applied to quasi-identifying attributes. A score close to 1 (or 100%) indicates minimal generalization.	Good for quick, simple assessment when data is primarily generalized using hierarchies (e.g., ZIP Code generalized from 12345 to 123xx).
Average Equivalence Class Size	Record-Oriented	Calculates the average size of the indistinguishable groups (\$k\$-groups). Larger class size implies higher privacy but lower utility.	Used as a simple proxy for k-anonymity's effectiveness; analysts may prefer

			groups that are not excessively large.
Workload-Aware Models	Special-Purpose	Measures utility by assessing the suitability of the output data for a specific machine learning task, such as training a classification model (e.g., Logistic Regression or Random Forest).	Used when the primary objective is to use the anonymized data for training a predictive model.

7. Workflow Summary

8. Load data into ARX.
9. Classify attributes (identifiers, quasi-identifiers, sensitive, insensitive).
10. Define hierarchies for generalization.
11. Choose privacy models (k-anonymity + extensions).
12. Run anonymization and evaluate risks.
13. Compare transformations using utility metrics.
14. Export anonymized dataset.

15. Input/output formats: description of the input/output of the tool at the validation stage (will help if some adaptation is needed).

Input: A tabular dataset in .csv or .xlsx

Output: 1. PDF report, describing the generalization level per attribute

2. CSV or excel file containing the modified clinical data or imaging meta-data

16. Quantitative results: performance obtained during training of the tool (if applies)

Not applicable.

17. Qualitative results: Provide some visual results (if available) of applying such tools.

Data Import

Each cohort (imaging metadata or clinical information) has to be structured in a tabular form in order to initiate a transformation, which is saved as a separate project. Example of tabular data prepared for import:

A	B	C	D	E	F	G	H	N	O	P	Q	R	S	T	U	V
1.3.6.1.4.11.3.6.1.4.1.58108.2023.53710477215516375985684415697707039957								078Y	103824	104212		104212	76		104212	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.165678093053841234019701522667484512329								064Y	125510	125659		125659	92		125659	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.243209474587198328091527157970750366182								064Y	125510	125620		125620	92		125619	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.162588903381932101931849865625510351362								064Y	125510	125511		125511	92		125510	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.214595635736140263813367723512853887755								069Y	192141	192141		192141	70		192141	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.19403594548818874614330271787611422822								069Y	192141	193212		193212	70		193212	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.258916596057617557451039694615462569215								069Y	192141	194030		194030	70		194030	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.21639183011841109089467111757520693709								069Y	192141	192757		192757	70		192757	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.284072469582509975456312146214705323284								069Y	192141	193147		192939	70		192141	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.163106801253695934532550597428377361959								069Y	192141	193658		193658	70		193658	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.287959416813001507282245040416135933								069Y	192141	194128		194128	70		194127	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.261241683805938690805264390902305792925								069Y	192141	192228		192228	70		192227	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.151422533422016151423606254488201492328								069Y	192141	193811		193811	70		193811	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.47508426656106497990108604320544845838								069Y	192141	193146		192939	70		192141	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.29612209523633008166449168786263632853								069Y	192141	194559		194559	70		194559	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.196832136627399180255873163718597448542								069Y	192141	192939		192939	70		192939	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.210873573172796019199707099281314242550								069Y	192141	192459		192459	70		192459	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.14072446698617771033531975430616853182								069Y	192141	192321		192321	70		192321	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.15964361057327573497681529467095380465								056Y	200542	201905		201905	72		201905	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.273175817724630038115136620145664350890								056Y	200542	201719		201719	72		201719	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.281877662863432390700367908999087184801								056Y	200542	200910		200910	72		200910	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.141367297492061910285028089948289101313								056Y	200542	202124		202124	72		202124	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.32176818173683592356855063167232135346								056Y	200542	202320		202320	72		202320	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.338751409611132602073766973181019060198								056Y	200542	201510		201510	72		201510	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.13245269639762251201753072774311061851								056Y	200542	201419		201419	72		201419	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.9153941987286473298720732066249140768								056Y	200542	200542		200542	72		200542	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.79880956909361851516411744139197986810								056Y	200542	202124		202124	72		202124	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.246861689210096607676023008820198563270								056Y	200542	200658		200658	72		200658	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.251271154683976880185745576239582165								056Y	123511	124323		124323	95		124323	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.860160651510679556241863516106626396								056Y	123511	124821		124821	95		124821	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.118133863710983507331579705718653011864								056Y	123511	125545		125545	95		125545	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.270160698504975240349328643980149027720								056Y	123511	124634		124634	95		124634	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.331426997066874221519784078562519800928								056Y	123511	125947		125947	95		125947	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.3216010273942564435779557224232480016								056Y	123511	125256		125256	95		125256	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.189107450825488205196771378561936302562								056Y	123511	123543		123543	95		123542	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.299755762411140843157919651715372781158								056Y	123511	124012		124012	95		124012	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.1041414761957932585303981811663323810								056Y	123511	130554		130554	95		130554	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.34178928003918609833955452056876126624								056Y	123511	123734		123734	95		123734	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.267791153158225337620758518964546359017								056Y	123511	125721		125721	95		125721	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.3182841355445946011008799728721635195								056Y	123511	125118		125118	95		125118	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.236261266128536860981437113039941124090								053Y	181814	182130		182130	90		182130	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.85673318649028143058073235953382305116								053Y	181814	181814		181814	90		181814	
1.3.6.1.4.11.3.6.1.4.1.58108.2023.4164054995660369581896580658132847600								053Y	181814	182623		182623	90		182623	

Figure 1: Example of DICOM metadata input in a tabular form. Each column contains a specific DICOM tag, while each row contains one case (one time-point of diagnosis/follow up imaging session of one patient). In this csv or excel format it can be used as input for the WIZARD tool.

Creating Hierarchies for the QIs

The next step is to identify the columns containing information that can be related to patient re-identification and then to be assigned the QI label instead of the default

Inensitive label. Then the user is allowed to import or create hierarchies to escalate the degree of information granularity to be considered when exploring the possible solutions for safe and usable outputs.

Data transformation		Attribute metadata	
Type:	Quasi-identifying	Minimum:	All
Level-0	Level-1	Level-2	Level-3
19960121	199601**	1996****	*****
20000107	200001**	2000****	*****
20000319	200003**	2000****	*****
20001126	200011**	2000****	*****
20010104	200101**	2001****	*****
20011026	200110**	2001****	*****
20011226	200112**	2001****	*****
20020929	200209**	2002****	*****
20021001	200210**	2002****	*****
20030716	200307**	2003****	*****
20031213	200312**	2003****	*****

Figure 2. Example of creating a date hierarchy by masking digits in a YYYYMMDD format.

Selecting an anonymization model

The user is guided through the interface to choose the most appropriate data modification scheme. Moreover suppression and generalization weighting are also modifiable and the user can refine the orientation regarding the proposed solution space, i.e. the elimination of more than a certain percentage of outliers can be restricted to a certain percentage of the population, in order to maintain usability. Moreover the attributes can have different weights when optimizing the proposed solution space, i.e weights can be unequal among attributes depending on their effect on the usability aspect.

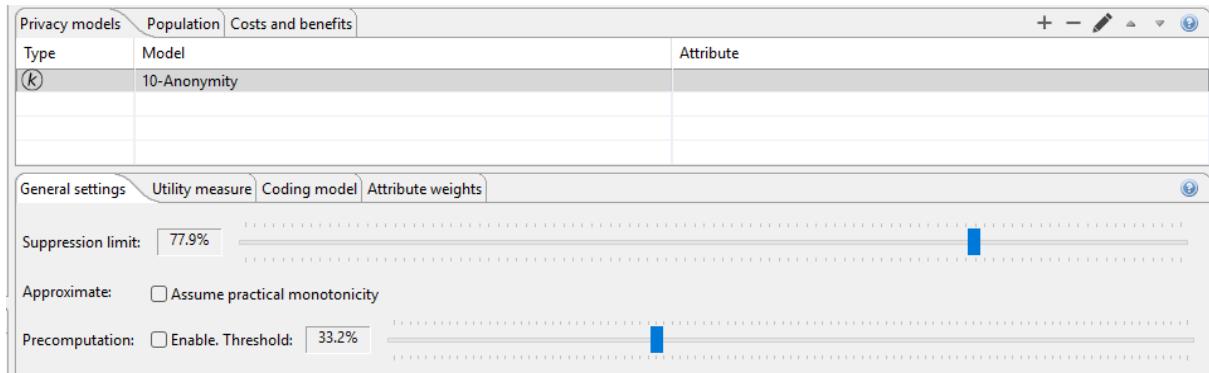


Figure 3. Example of anonymity model configuration. In this figure k-10 anonymity is selected. The allowed suppression limit is set to constraint the percentage of outliers that can be ignored for the analysis.

Running the analysis

Running the analysis will create in the “Explore Results” tab a graphical overview of the possible solutions, i.e the modifications in attribute granularity that fulfill the requirements defined regarding and usability during the previous steps. The user is able to define filters in the proposed configurations in order to simplify the graph by presenting only the subset of solutions meeting a certain constraint or a specification, i.e. excluding a certain hierarchy from a certain attribute, thus reducing the available possible configurations.

Colour coding denotes the compliance with requirements, with yellow representing the optimal in the category of generalization, and green the rest of available options. The shape shows selected / unselected solutions. Once a choice is made by the user, by right clicking the configuration is applied and the output is presented in the analysis of utility in the appropriate tab.

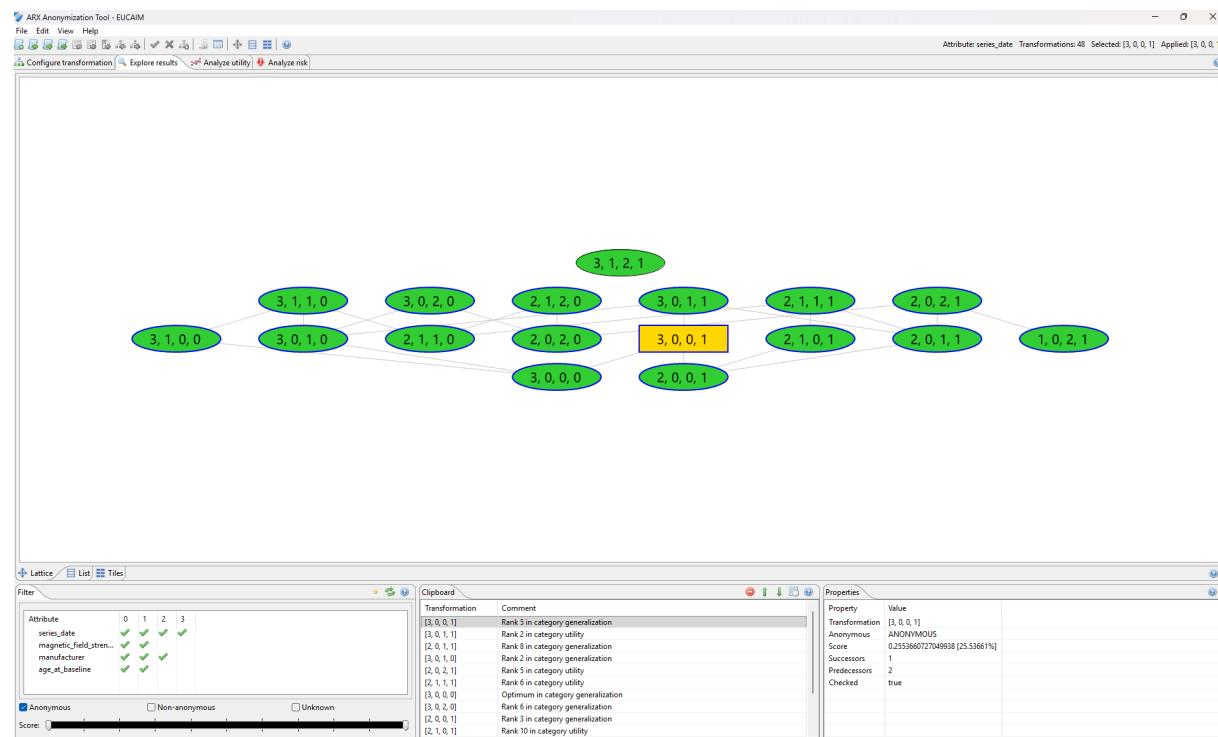


Figure 4. The exploration of the solution space shown in a graphical form. Each shape contains a possible solution with each vector showing the chosen level as defined in the hierarchies for each QI. In the table at the lower middle of the screen the ranking of each solution presented in the solution space is ranked with respect to security and usability criteria. The user is allowed to select and apply any combination (presented as vector) of generalization. For any such combination, metrics for the transformers attributes are calculated and presented.

EUCAIM WP5 Task Force 2

Tools for data quality assessment and data cleaning

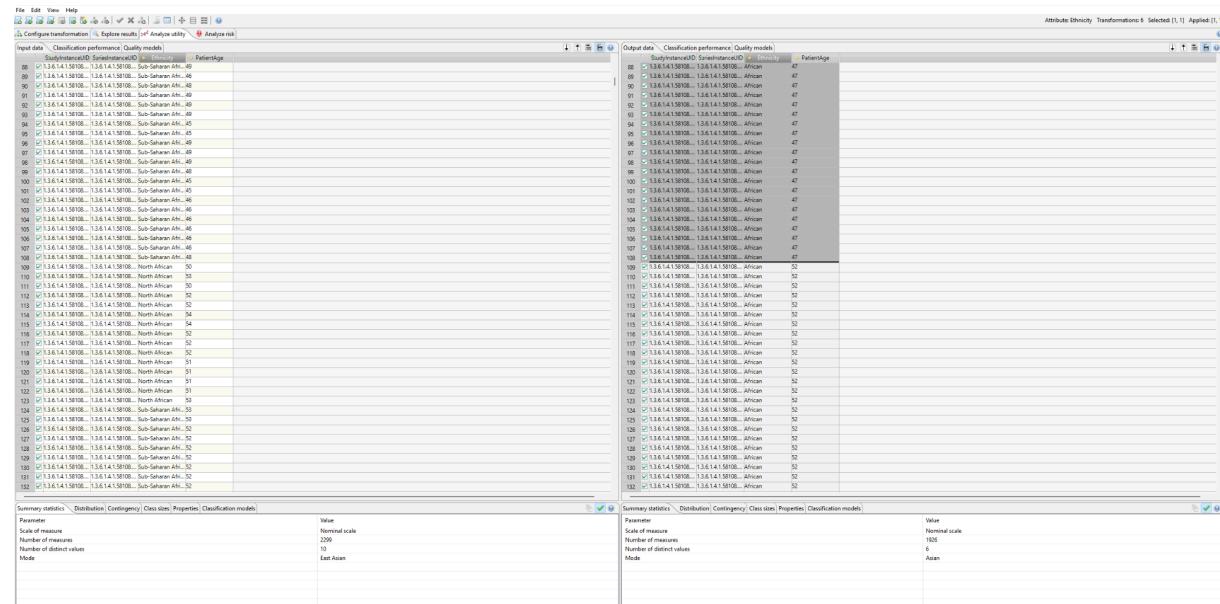


Figure 5. Example of the original (left) and modified (right) data after applying the modification schema. Grey colour on the right side denotes equivalent cases.

After applying the modifications, a utility analysis step will help the user redefine the solution and conclude to the final optimal solution, chosen to maintain anonymity and also data value for deployment under certain tests and requirements.

Utility analysis

Preserving data utility can be a contradictory notion of security as depriving data of patient specific information can compromise usability, in certain cases. The degree of effect differs among working scenarios: summary statistics, counts, machine learning, or ad-hoc analysis will tolerate different kinds and amounts of distortion.

- Generic information-loss metrics: quantify how much data was generalized or suppressed overall (attribute-level loss).
- Equivalence-class metrics: summarize how records were grouped (e.g., average or distribution of equivalence class sizes).
- Result-based (task-specific) utility: measure utility by running the actual analysis you care about (e.g., train a classifier on anonymized data and compare accuracy with the original).
- Distributional and visual checks: compare histograms, counts and value distributions between original and anonymized data to find large distortions or bias.

EUCAIM WP5 Task Force 2

Tools for data quality assessment and data cleaning

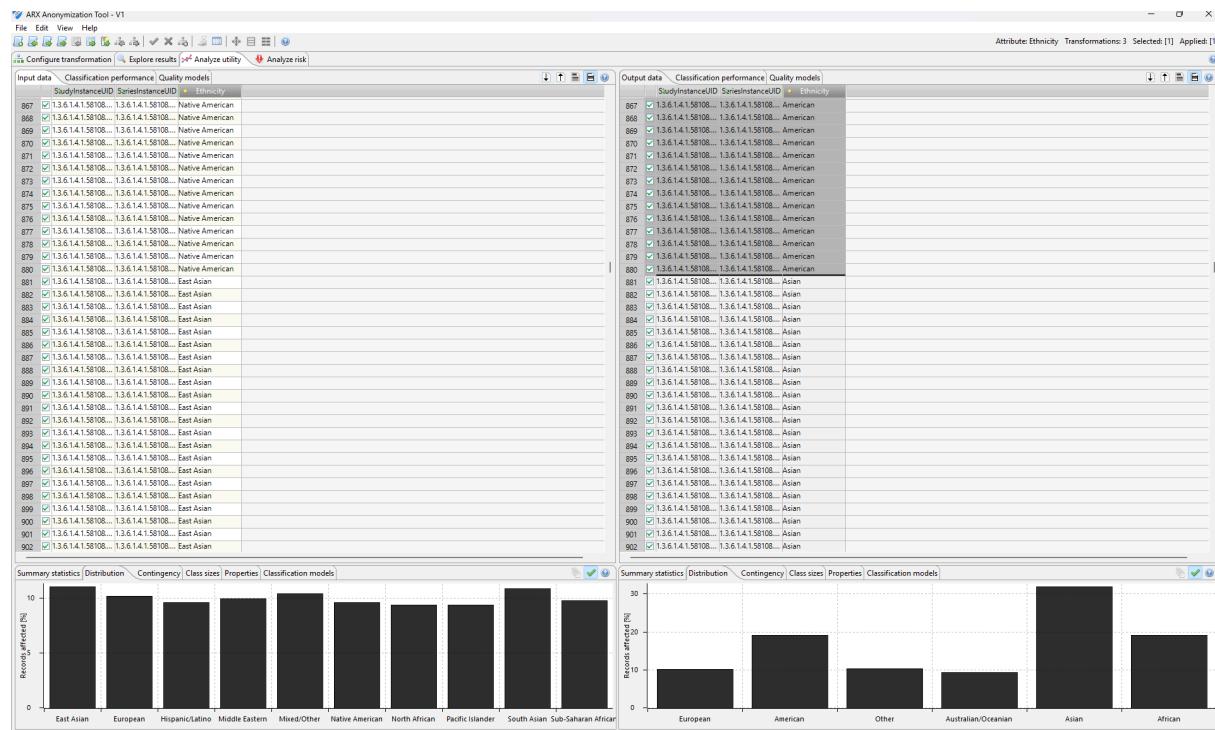


Figure 6. Graphical representation of the output, corresponding to the selected solution. Each class population, before (left) and after (right) applying the selected solution is an indication of the generalization degree for the selected QI.

Contingency plans

In ARX, a contingency plan usually refers to steps prepared in case the anonymization process fails to meet privacy constraints, data utility thresholds, transformation limits. In this context, a contingency plan could involve:

- Choosing alternative generalization hierarchies
- Relaxing privacy constraints
- Switching to suppression instead of generalization
- Using different risk models
- Preparing alternative anonymization strategies
- Falling back to coarser anonymization to guarantee compliance

- The colour coding graph presents original and anonymized data as seen in Figure X.

The colors show how aggressive you're allowed to be within the allowed minimum and maximum generalisation/suppression limits. Colors usually mean:

Green – Safe / Minimal Transformation

- The attribute will be changed *as little as possible*.
- Low generalization or low suppression in contingency mode.

Yellow – Moderate Transformation

- ARX is allowed to generalize or suppress this attribute more than usual.
- A middle-ground fallback option.

Red – Maximum Transformation Allowed

- The attribute may be heavily generalized or even heavily suppressed in the contingency plan.
- Utility loss may be high.

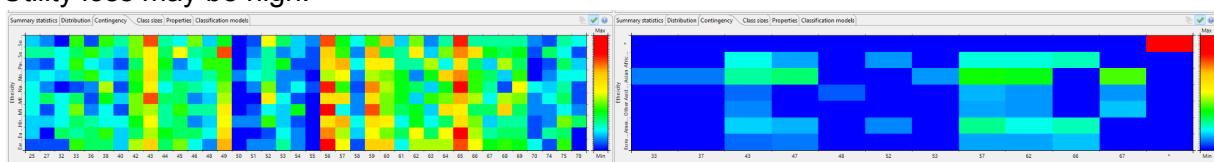


Figure 7: Example of Contingency plan, before (left) and after (right) applying the selected solution.

Classification Models

In ARX, “classification models” are machine-learning classifiers used to evaluate how well an anonymized dataset can still predict a chosen target attribute (the “class attribute”). They measure data utility loss by comparing: Performance of the classifier on the original data vs Performance on the anonymized data

ARX provides several ML models. Each model reacts **differently** to anonymization, i.e. Naive Bayes checks distribution preservation, logistic regression checks linear relationships, Decision trees checks structural patterns, Random forest for overall predictive capacity and lastly SVM checks for boundary preservation. Using multiple classifiers gives you a **robust picture** of how much predictive utility you retained. This helps to define whether anonymization is too strict, how much predictive power the anonymized dataset still has and whether alternative hierarchies or constraints improve utility

EUCAIM WP5 Task Force 2

Tools for data quality assessment and data cleaning

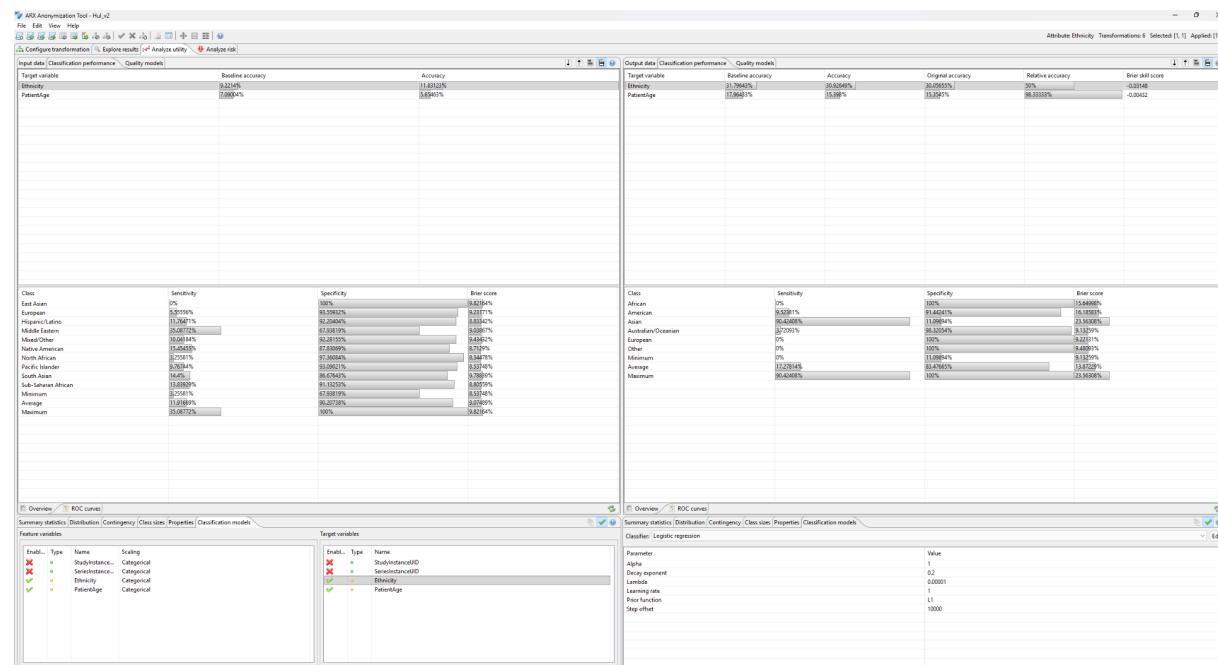


Figure 8. Utility analysis before (left) and after (right) applying the selected solution, by quantitative analysis of utility metrics.

Hierarchies

Importing or creating hierarchies involves the feeding the generalization capabilities in the form of a csv or excel file, registering the different granularities for each attribute. Predefined csv files can be found in ???, but can also be created by the users based on their specific input and cohort requirements.

B	C	D
Level 0 -Country	Level 1 -Continent	Level 2
Africa	Africa	*
Nigeria	Africa	*
Egypt	Africa	*
South Africa	Africa	*
Asia	Asia	*
China	Asia	*
Japan	Asia	*
India	Asia	*
Europe	Europe	*
France	Europe	*
Germany	Europe	*
Greece	Europe	*
North America	North America	*
United States	North America	*
Canada	North America	*
Mexico	North America	*
South America	South America	*
Brazil	South America	*
Argentina	South America	*
Colombia	South America	*
Oceania	Oceania	*
Australia	Oceania	*
New Zealand	Oceania	*

Figure 9. Example of csv file for applying a 3-level hierarchy for nationality.

18. Additional information: successful use cases, external resources (open code, papers...) licence, certification ... (if they apply).

An indicative list of publications related to successful use cases is presented next.

The core publication of the tool (with successful use cases):

DOI: 10.1002/spe.2812

Flexible data anonymization using ARX—Current status and challenges ahead

Prasser F. Eicher J. Spengler H. Bild R. Kuhn K.A.

Software - Practice and Experience, 2020

DOI: 10.1038/s41597-020-00773-y | PMID: 33303746

Cited by

| PMCID: PMC7729909

Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19

Jakob C.E.M. Kohlmayer F. Meurers T. Vehreschild J.J. Prasser F. Scientific Data, 2020

○

Keywords for searching in databases: anonymization tool

III. Technical specifications

This part of the documentation is dedicated to providing all relevant technical specifications to prepare for the tool's integration into the EUCAIM test environment.

It is possible that the tool's integration will require, among others, some modifications in the input/output, or the inclusion of monitoring mechanisms.

1. Data: In depth description of the data used to train the tool.

The tool is not an AI tool. It has been tested on clinical data across various diseases including cardiovascular diseases, autoimmune diseases, mental disorders.

2. Methods: in depth description of the methodology used for its development including all data preprocessing.

The ARX software library comes in two flavours: (1) libarx, which contains all dependencies required for using all of ARX's features and, (2) libarx-min, which doesn't come with any external libraries included. This page lists all libraries that must be included together with libarx-min to use certain features of ARX.

Minimal requirements

To use the basic features of ARX, the following libraries must be included:

- [Colt](#): Open Source Libraries for High Performance Scientific and Technical Computing. The version currently used by ARX can be found [here](#).
- [HPPC](#): High Performance Primitive Collections for Java. The version currently used by ARX can be found [here](#).
- [Commons math](#): The Apache Commons Mathematics Library. The version currently used by ARX can be found [here](#).
- [JHPL](#): Java High-Performance Library for Lattices. The version currently used by ARX can be found [here](#).

Risk analyses and risk-based anonymization

- [Newton-Raphson](#): Java implementation of Newton's method for solving bivariate non-linear equation systems. The version currently used by ARX can be found [here](#).
- [Commons validator](#): Apache Commons library for verifying the integrity of data. The version currently used by ARX can be found [here](#).

Utility analyses using classification models

- [exp4j](#): For evaluating expressions and functions. The version currently used by ARX can be found [here](#).
- [Apache Mahout](#) and dependencies: To add support for some machine learning algorithms. The version currently used by ARX can be found [here](#).
- [SMILE](#): To add support for some machine learning algorithms. The version currently used by ARX can be found [here](#).

Importing data

The following libraries must be included for importing data from certain sources:

- [Univocity CSV Parser](#): Text parsing solution for Java that provides a common architecture for parsing tabular representations of data. The version currently used by ARX can be found [here](#).
- [Apache POI](#): Java API for Microsoft Documents. This library is required for reading data from Excel Spreadsheets. The version currently used by ARX (and its respective dependencies) can be found [here](#).
- JDBC drivers: ARX currently supports [MySQL](#), [PostgreSQL](#), [MS SQL Server](#), [Oracle](#) and [SQLite](#). The versions currently used by ARX can be found [here](#), [here](#), [here](#), [here](#) and [here](#).
- [Apache Commons IO](#): Library of utilities to assist with developing IO functionality. This library is required for using the advanced importing features for CSV files (via class DataSource). The version currently used by ARX can be found [here](#).
- Selecting data subsets

To select data subsets (e.g. for enforcing δ-presence) by means of a querying interface, the following library must be included:

- **Object-selector:** Object Selection Library for Java. The version currently used by ARX can be found [here](#).

3. Specific Technical information:

- CPU:**
- Programming language:**
- Expected RAM usage:**
- Running mode:** case-based.
- Software version:** v1.0.0.
- Libraries:**
- Security measures:**

4. Traceability and monitoring mechanism.

Currently no mechanism for traceability and monitoring

5. Unitary tests: description of the tests implemented to verify the correct functioning of the tool.

6. Access restriction : do you have any access restriction to the source code or to the binaries of the tool?

No, the tool is based on the open source ARX software, available under no requirements.

7. Additional information for tool integration.

Continuous optimization of the provided algorithm and toolset is performed.

IV. Integration Validation

In this stage, further documentation is required by the tool providers. In particular, the following important points are suggested to be described about the tools:

1. Communication channel for the helpdesk, technical support channel:

We have a team of developers who can support any errors regarding the tool.

2. Most common errors:

Uploading invalid input dataset format. Too strict anonymity requirements



FAQs:

Currently there is not a list of FAQs.

3. User Manual:

a. Installation/configuration instructions

<https://arx.deidentifier.org/anonymization-tool/>

b. Download software from :

<https://arx.deidentifier.org/downloads/>

: