Data Anonymization Project

Enseignant

- Amin EHSAN

TEAM 4

- DIANGUE DI MOUNGUET
- Ousmane MOMBO MOMBO
- Cassandre NONGO
- Bally-Stone MOUDIANGO

INSTRUCTIONS

The main objective is to learn how to protect the privacy of individuals present in a real-world dataset, while preserving the analytical value of the information.

The dataset used comes from a medical study on AIDS, referred to here as the AIDS dataset. It contains detailed information on patients monitored in a clinical setting: age, gender, Karnofsky score (assessing the patient's general condition), CD4 count (an immune indicator), as well as behavioral and treatment variables.

This type of data is particularly sensitive because it combines confidential medical information and personal characteristics. Mismanagement or dissemination of this data could lead to the reidentification of individuals or even a violation of their medical confidentiality.
It is in this context that anonymization becomes essential.

The objective of this project is twofold:

1. To reduce the risk of re-identification of individuals in the database using several anonymization methods derived from Statistical Disclosure Control (SDC).

2. To assess the impact of these methods on data quality and structure to ensure that statistical analyses remain relevant.

The entire process is based on a central principle: finding the right balance between data confidentiality and utility.

In other words, we seek to protect patient privacy without rendering the dataset unusable for medical research.

Project organization

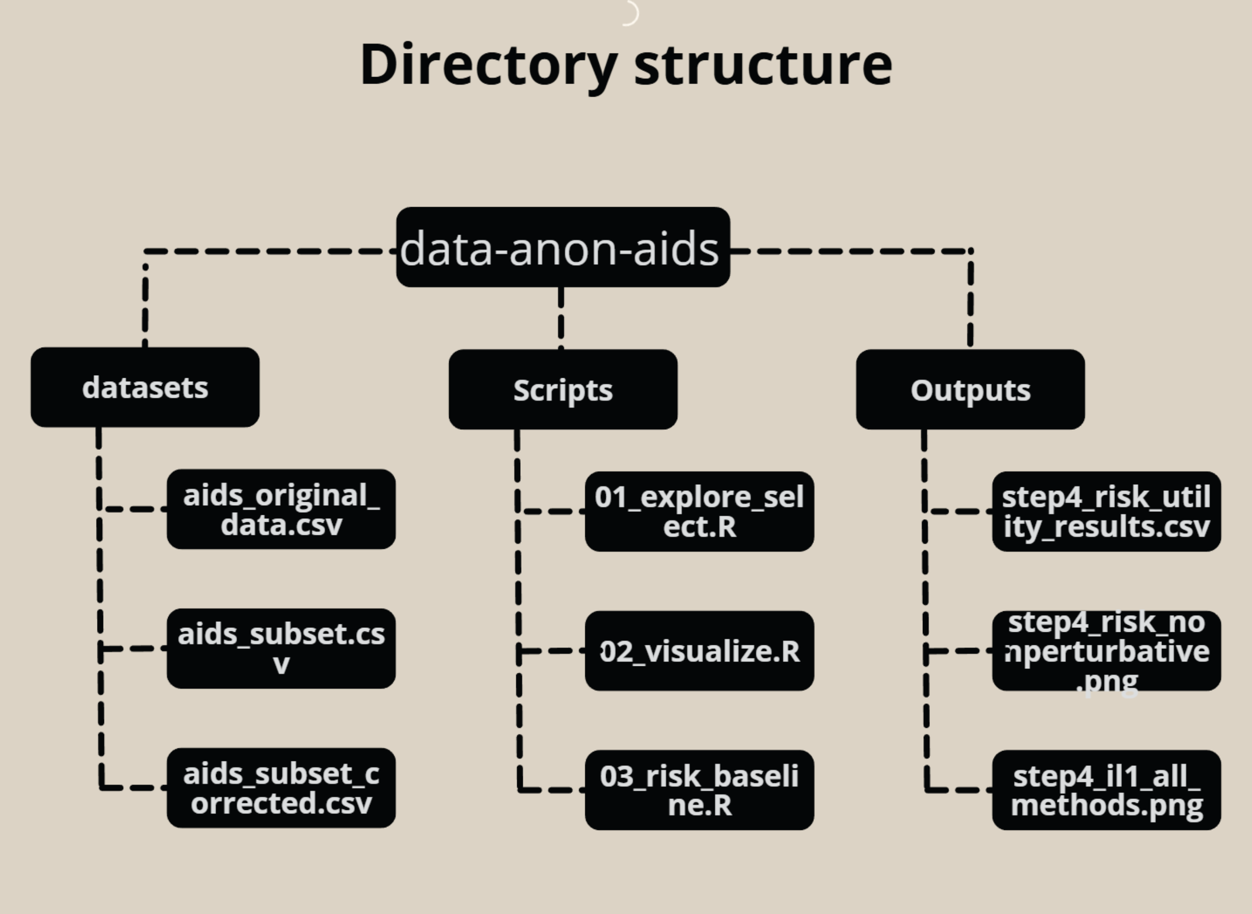
The project is organized in a modular manner, with several stand-alone R scripts, each corresponding to a step in the anonymization process:

01_explore_select.R: Exploration and selection of relevant variables.

02_visualize.R: Visualization and understanding of data distribution.

03_risk_baseline.R: Application of anonymization methods and risk assessment.

The generated results, graphs, and CSV files are saved in the outputs/ folder. This structure allows for complete reproducibility of the project and a clear separation of analytical steps.



Loading and exploring

- The original dataset was imported, verified, and then cleaned (aids_original_data.csv).

- The variables were categorized into:

Quasi-identifiers: age, gender, race, homosexual, drugs

Sensitive variables: cd40, karnof

Treatment/outcome variables: treat, symptom, days, cens

A relevant subset was created (aids_subset.csv).

Rationale: This subset preserves the essential diversity of the data while limiting the complexity of the variables to be anonymized.

Variable selection and typing

Once the variables were selected, particular attention was paid to their typing.
Binary variables (e.g., gender, homosexual, drugs, etc.) were converted into factors but retained their numeric labels (0/1).

This choice is crucial in an anonymization context: retaining explicit labels (such as “Male/Female” or “Yes/No”) could facilitate re-identification or introduce semantic bias.
Continuous variables (age, cd40, karnof, days) were left in numeric format to allow for subsequent statistical analyses.

Finally, a derived variable, age_band_10y, was created to group ages into 10-year bands.
This overall recoding paves the way for methods to reduce the risk of re-identification.

Exploratory visualization

Before applying the anonymization methods, an exploratory visualization phase was conducted to better understand the structure of the dataset.

Histograms were used to observe the distribution of continuous variables (age, CD40, days), while bar charts illustrated the distribution of categorical variables (gender, homosexual, race, drugs).

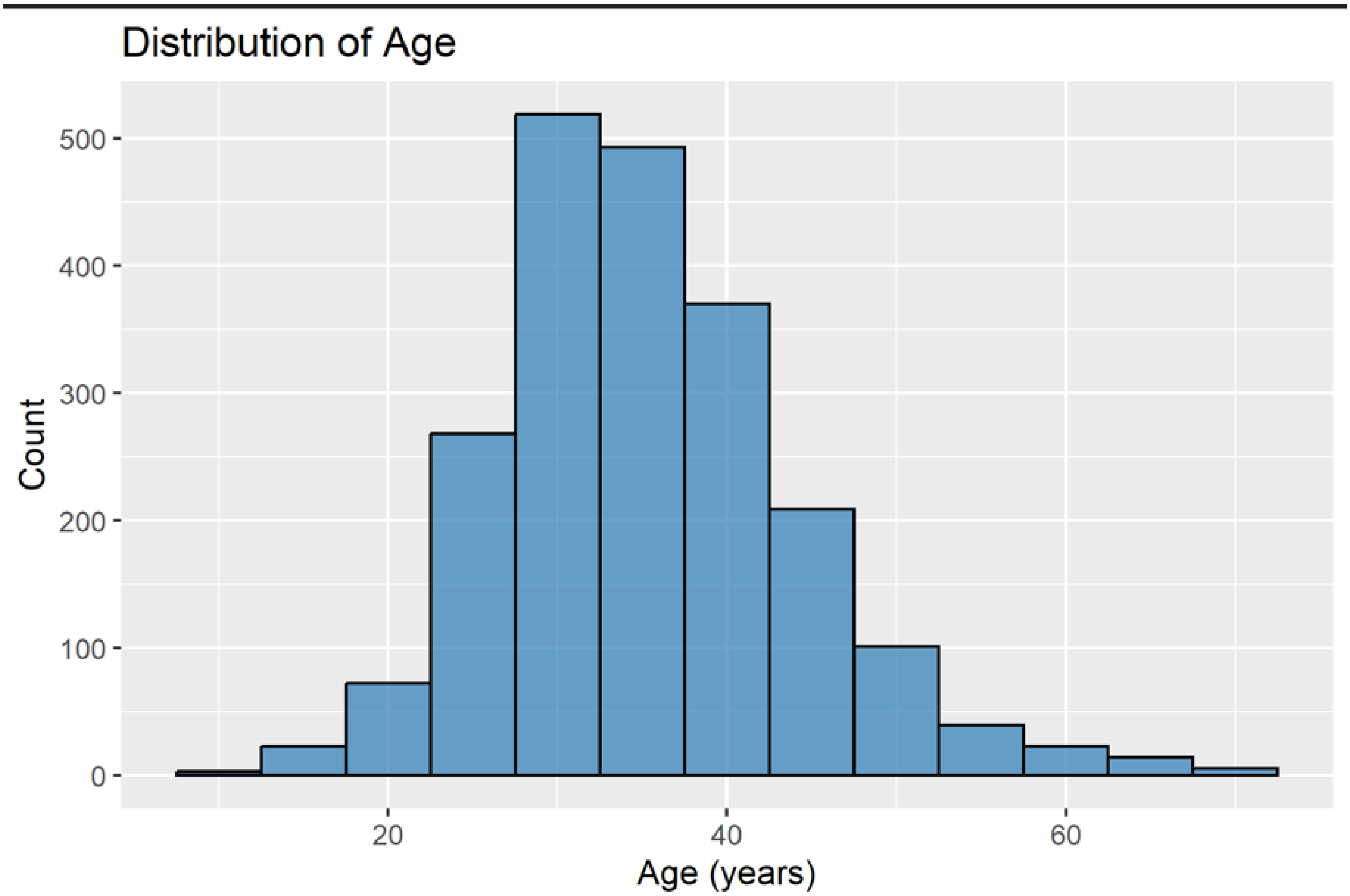
These visualizations highlighted several important elements:

- A high concentration of ages between 30 and 50.
- A skewed distribution of CD4 counts, with some outliers.
- An imbalance in certain binary categories (for example, the majority of individuals are male and homosexual).

These observations indicate that certain combinations of variables can be nearly unique, and therefore particularly risky from a re-identification perspective.
They guided the selection of the most appropriate anonymization methods.

Age Histogram

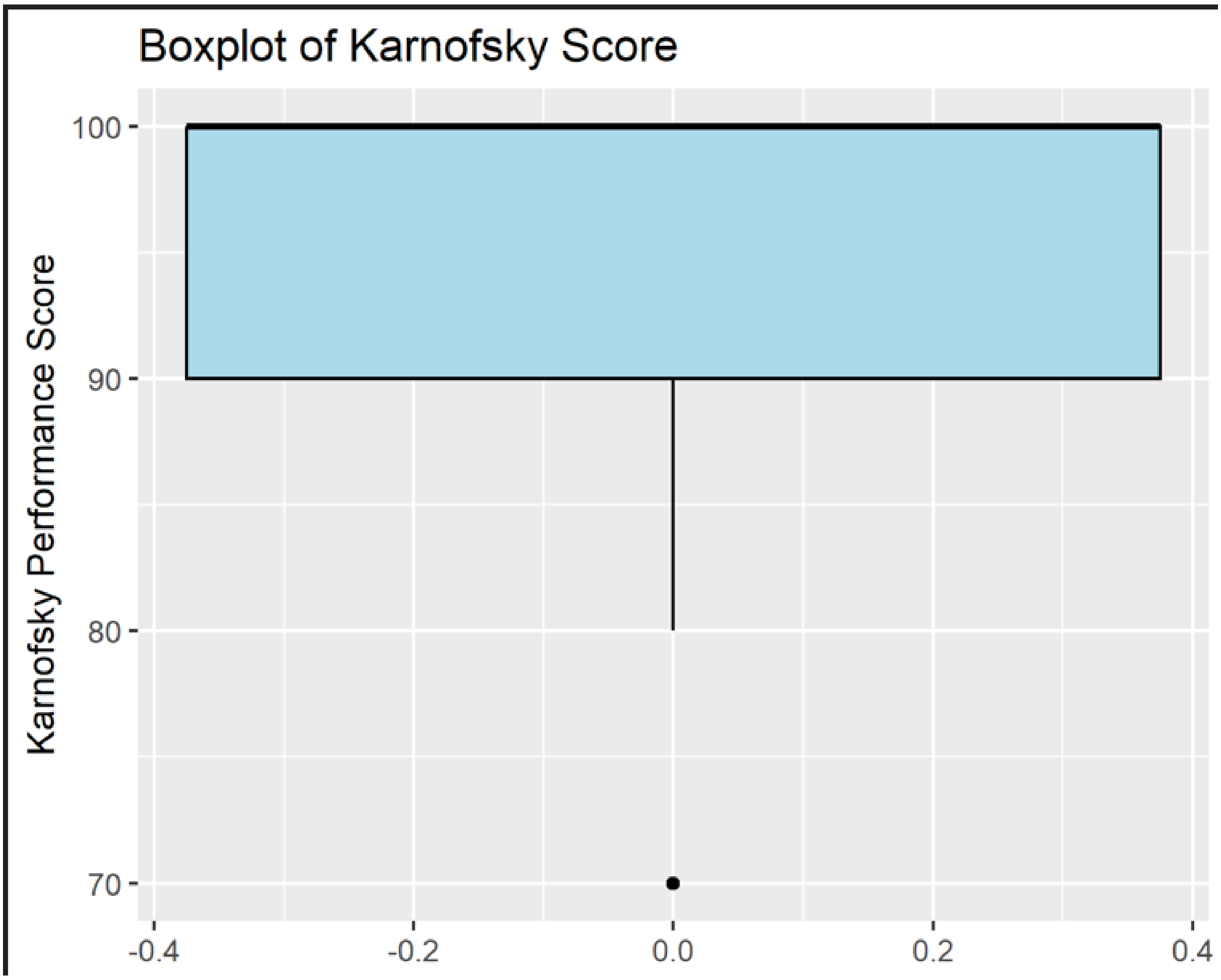
Objective: To visualize the distribution of ages to understand whether certain age groups are rare (risk of uniqueness).



The age distribution shows a concentration between 30 and 50 years, which indicates an active adult population. The extreme age groups (<25 and >60 years) are poorly represented and may increase the risk of re-identification.

Karnofsky Score Boxplot

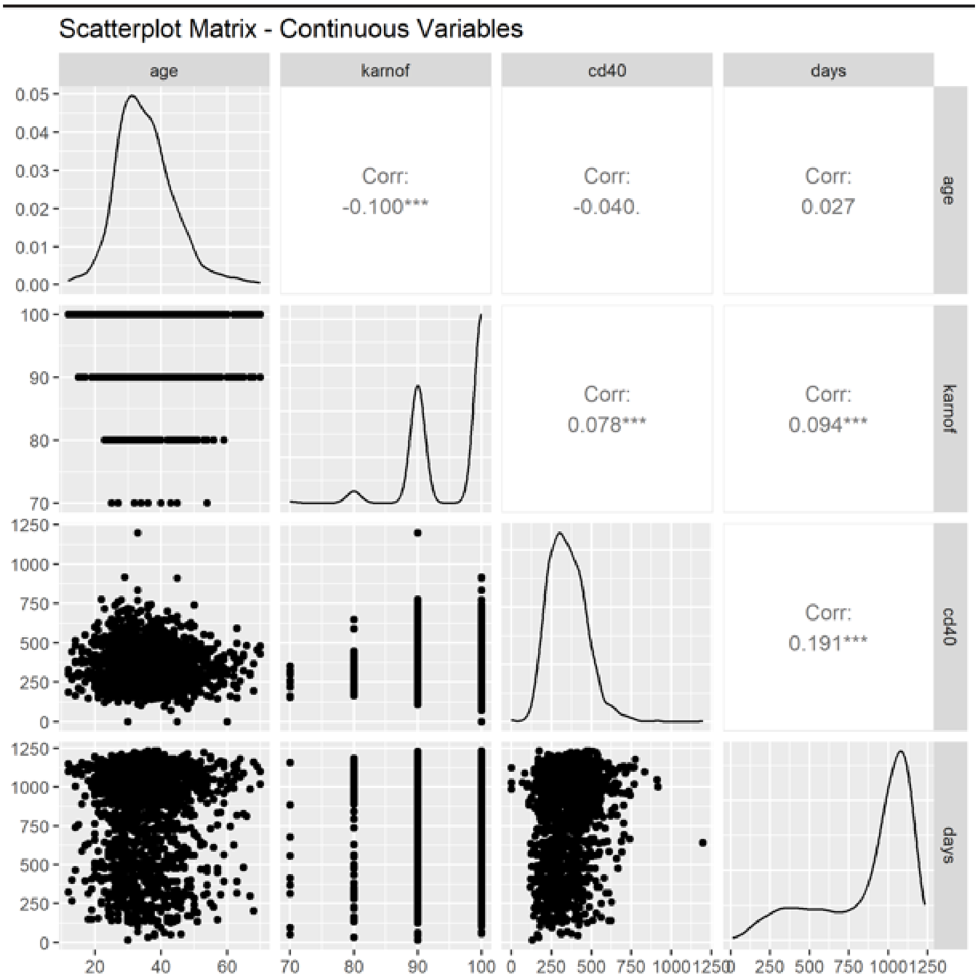
Objective: To detect extreme values and variability in patients' clinical status.



The Karnofsky score is strongly centered around 90–100, which indicates a relatively stable general condition among patients. Little dispersion means little diversity for this variable, thus a limited risk of reidentification.

Correlation matrix (or scatterplot matrix)

Objective: To study the relationships between continuous variables (age, Karnof, CD40, days).



The correlation matrix reveals a weak correlation between age and medical variables, but a moderate correlation between cd40 and karnof. This relationship should be preserved during anonymization to maintain the analytical consistency of the dataset.

Application of anonymization methods

Several approaches were tested to reduce the risk of re-identification while maintaining the usefulness of the dataset.

Non-perturbative methods

These methods modify the data structure without altering the values:

- Local Suppression (k-Anonymity): Removes or masks certain rare values to obtain at least k individuals sharing the same key combination.

→ Tested for k = 2, 3, 5, 10.

• Global Recoding: Groups continuous or categorical values into larger classes.

→ Applied to age with groupings into 3 and 2 bands.

Perturbative methods

These methods alter the values to reduce the risk of recognition:

• Microaggregation: Replaces continuous values with the average of small groups (groups of 3 and 5 individuals tested). • Additive Noise: Adds random noise to continuous variables (cd40, karnof, days) with an intensity of 5% and 10%.

• PRAM (Post Randomization Method): Randomly reassigns the categories of a variable (here race), with a probability of 5% and 10%.

Measurement and evaluation

a. Risk Measures

Two main indicators were used:

Global Risk: represents the average probability that an individual could be re-identified based on their quasi-identifiers.
In the original dataset, this risk was 0.0285 (2.85%), meaning that approximately 61 individuals out of 2,139 could be re-identified.
After anonymization, certain methods (recoding and deletion) reduced this risk to approximately 1.2%.

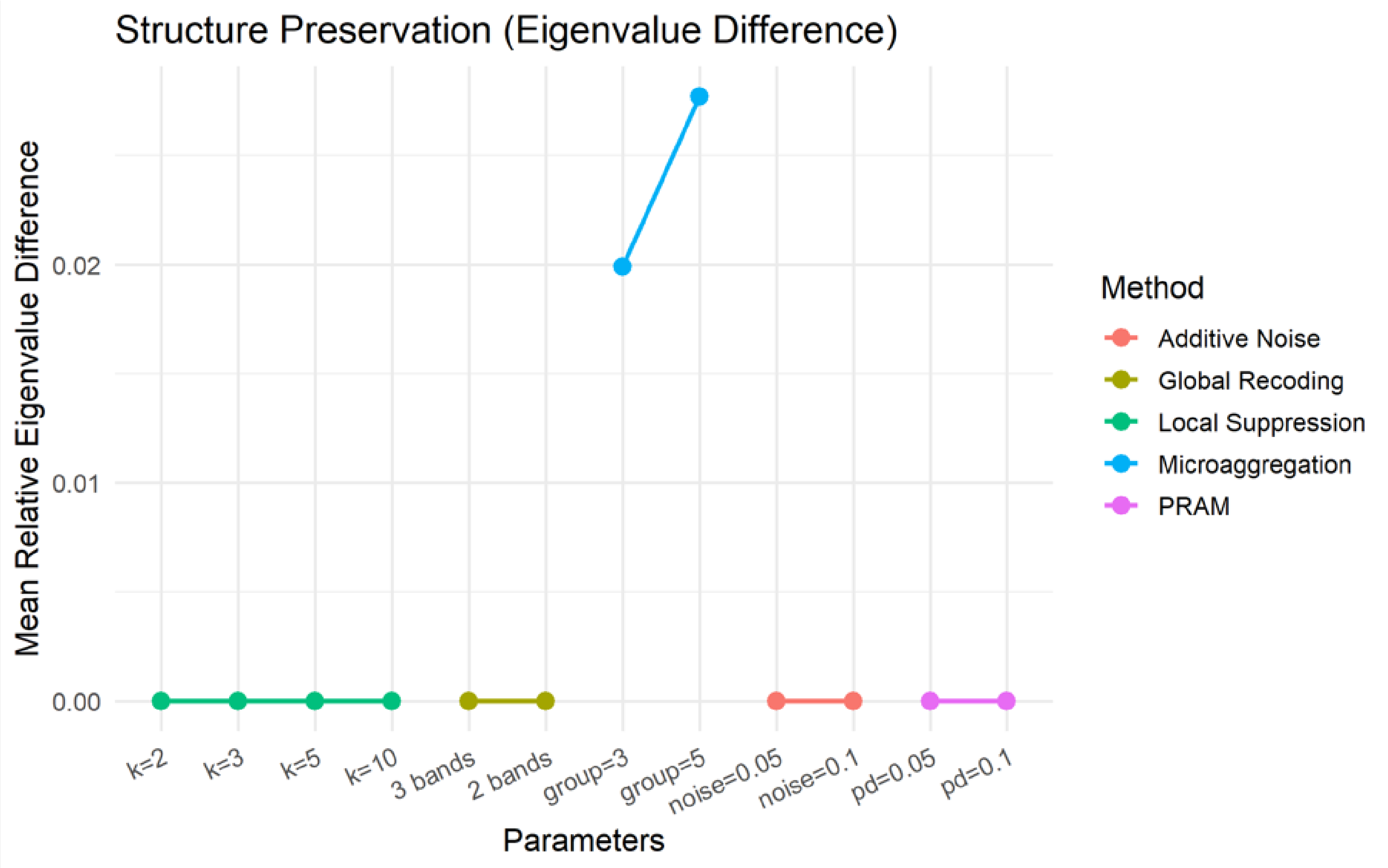
Expected Re-ID: corresponds to the estimated number of individuals that could be identified in the dataset.
This measure evolves along with the Global Risk: the stronger the clustering or the more numerous the deleted values, the more this value decreases.

b. Information Loss Measures

Two other measures were used to assess the impact of anonymization on data quality:
IL1 (Information Loss Index): measures the average difference between the original and modified values.
A low value indicates good data preservation. Non-perturbative methods maintain IL1 = 0, while microaggregation presents very high values, indicating significant information loss.

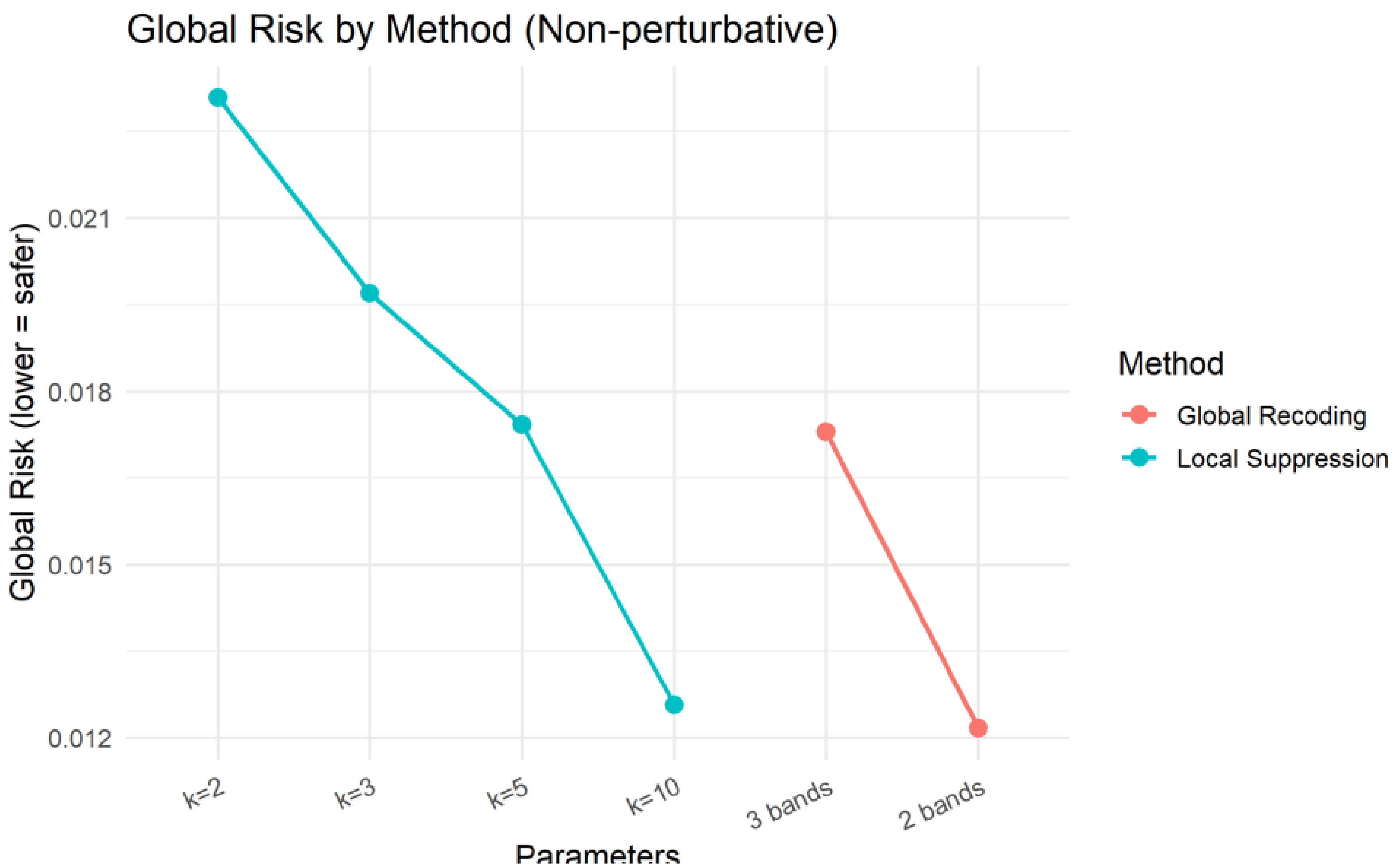
Eigenvalue Difference: compares the correlation structure of variables before and after anonymization.
A difference close to zero indicates that the overall structure of the data has remained stable. All methods, except microaggregation, maintained good structural consistency.

Graphique IL1 (Information Loss) (step4_il1_all_methods.png)



The IL1 graph illustrates the information loss associated with each method.
Non-perturbative methods (deletion, recoding) maintain maximum utility (IL1 = 0).
Microaggregation exhibits a very high IL1, indicating significant distortion of continuous data.
The addition of noise (5–10%) remains moderate and represents a good compromise between anonymization and accuracy.

Graphique du risque global (step4_risk_nonperturbative.png)



This graph shows the gradual decrease in overall risk as the k parameter increases (Local Suppression) or as age groups are grouped together (Global Recoding).
When k = 10 or the ages are merged into two bands, the risk falls below 0.02, which guarantees an acceptable level of confidentiality according to SDC standards.

Results and interpretation

Analysis of the results showed a significant reduction in the risk of re-identification, from 2.85% to approximately 1.2%, thanks to the combined methods of global recoding and local suppression.
Non-perturbative methods (suppression, recoding) proved to be effective and secure, while preserving data integrity.
Perturbative methods (microaggregation, noise) offered stronger protection, but at the cost of a measurable loss of information.

The resulting graphs illustrated these trends:
• Overall Risk Graph: Shows the decrease in risk as k increases or as age classes are merged.
• IL1 Graph (Information Loss): Highlights the degradation in accuracy for microaggregation and additive noise.
• Structure Graph (Eigenvalues): Shows that the statistical structure remains very stable for the majority of methods. In summary, the best performance was achieved with a combination of:

Global Recoding (2 bands) + Local Suppression (k=5) + Additive Noise (5%).
This combination provides an excellent balance between privacy and utility.

Conclusion

This project provided a practical application of the principles of data anonymization and an understanding of the trade-off between confidentiality and analytical utility.
The results demonstrate that it is possible to effectively protect individual privacy while maintaining scientifically usable data.

The approach adopted—exploration, visualization, anonymization, evaluation, and interpretation—represents a comprehensive professional approach.

The methods tested and the results obtained could serve as a model for the anonymization of real-world medical data.

In conclusion, this work illustrates how a rigorous anonymization process can transform a sensitive dataset into one that is secure, ethical, and still useful for research.