Anonymization Project report on AIDS data set

Elvis OBOUNOU, Sara EL KARDI, Wiam BELOUARD, Davis JOSEPH (Aivancity PGE-4 Grande école, 26-10-2025)

---

1) Executive summary

- We analyzed the aids_original_data.csv dataset (2,139 rows × 27 columns).
- **Step 1–2 (Exploration & Visualization):** We identified continuous, categorical, and sensitive variables; produced histograms/boxplots, categorical bar charts, pairwise plots and correlations.
- **Step 3 (Risk assessment in R):** Using sdcMicro, we quantified disclosure risk on quasi-identifiers **{age, gender, race}**. Baseline results:
  - **Global re-ID risk: 8.51%**
  - **Expected re-identifications: 182 / 2139**
  - **% unique (k = 1): 1.36%**
  - **% with k ≤ 5: 10.52%**
  - **Inference risk:** arms → treat is **deterministic** (attribute disclosure).
- **Step 4 (Anonymization in Python):** We varied parameters for:
  - **Age banding** (5, 10, 15 years)
  - **PRAM** on **gender** or **race** (flip p = 1%, 5%, 10%)
    Key outcomes:
  - **Age banding** dramatically reduces risk with modest utility loss.
    - 10-year bands: **expected re-IDs 25**, **k≤5 = 0.61%**, **IL1 ≈ 0.044**, **eigen-sim ≈ 99.75%**.
  - **PRAM** on a single binary key barely helps here; **age** is the dominant risk driver.
- **Recommendation:** Use **10-year age bands**, and **do not release both arms and treat** together (drop/merge one).

---

2) Dataset overview (Step 1)

Variables of interest

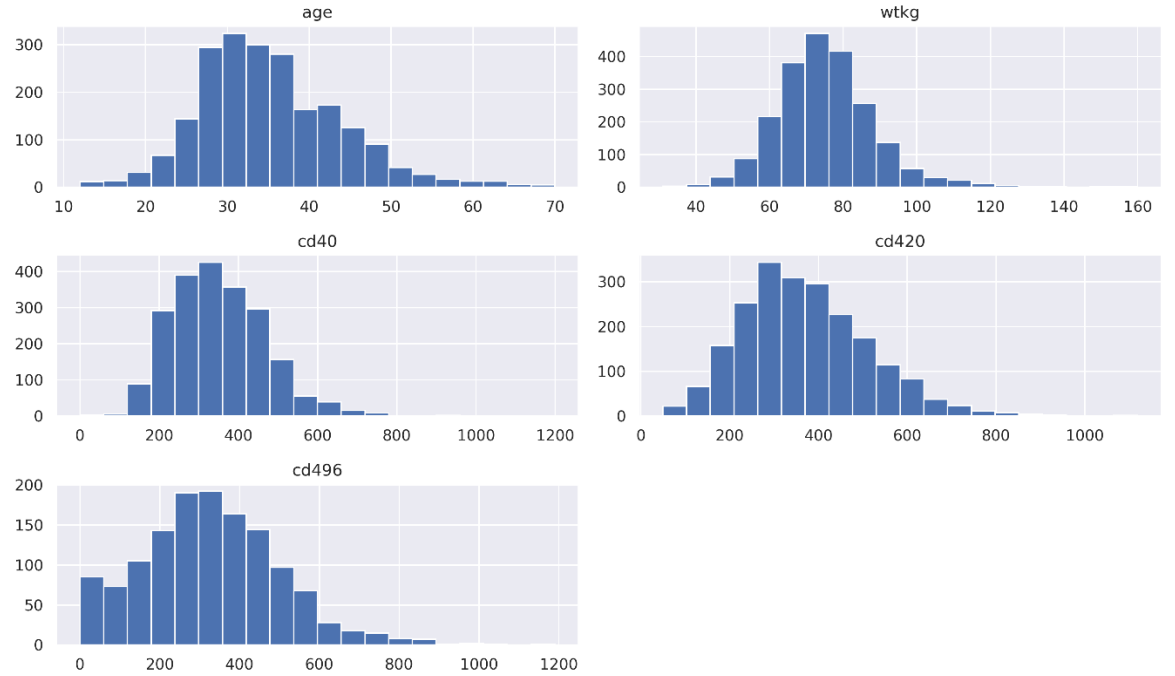
- **Continuous (examples):** age, wtkg, preanti, cd40, cd420, cd496 (≈37% missing), cd80, cd820, days, karnof.
- **Categorical / binary:** race, gender, homo, hemo, drugs, treat, arms (0–3).
- **Sensitive fields (for privacy):** race, gender, homo, hemo, drugs, and downstream attribute **treat**.

Notable data features

- zprior is constant (=1) → drop for modeling.
- arms perfectly indicates treat (0→0, 1–3→1).
- Outliers: high wtkg and high CD8 values; some CD4 zeros (likely true lows).

Figures

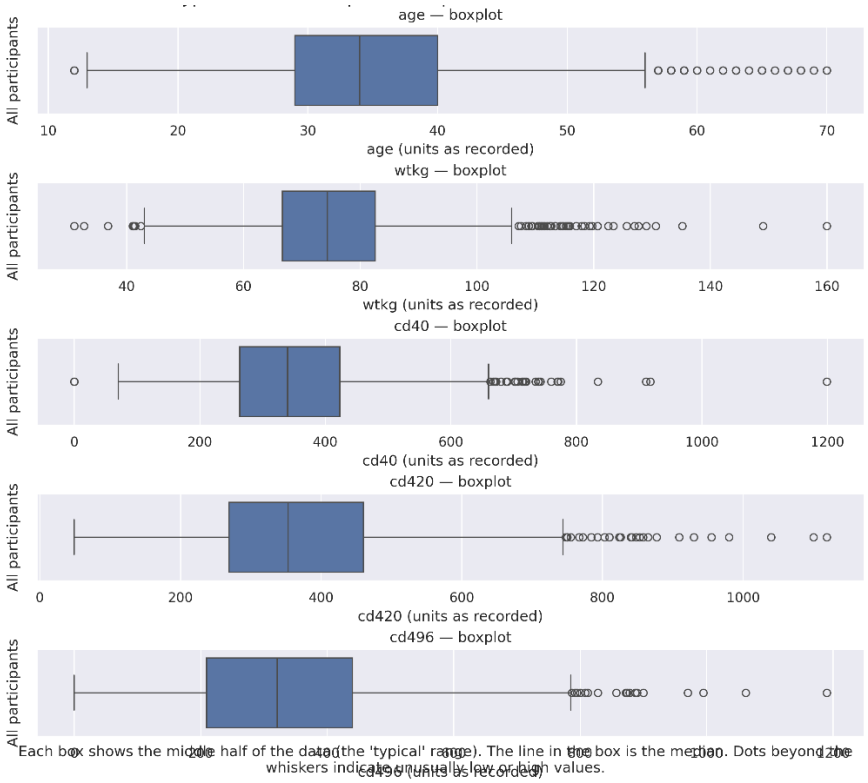
- **Figure S2-1:** Histograms — continuous variables



Each panel shows how many participants fall into each value range. This helps us see common values and whether there are very low or very high outliers.

step1_and_step2/outputs/figs/01_hist_continuous.png

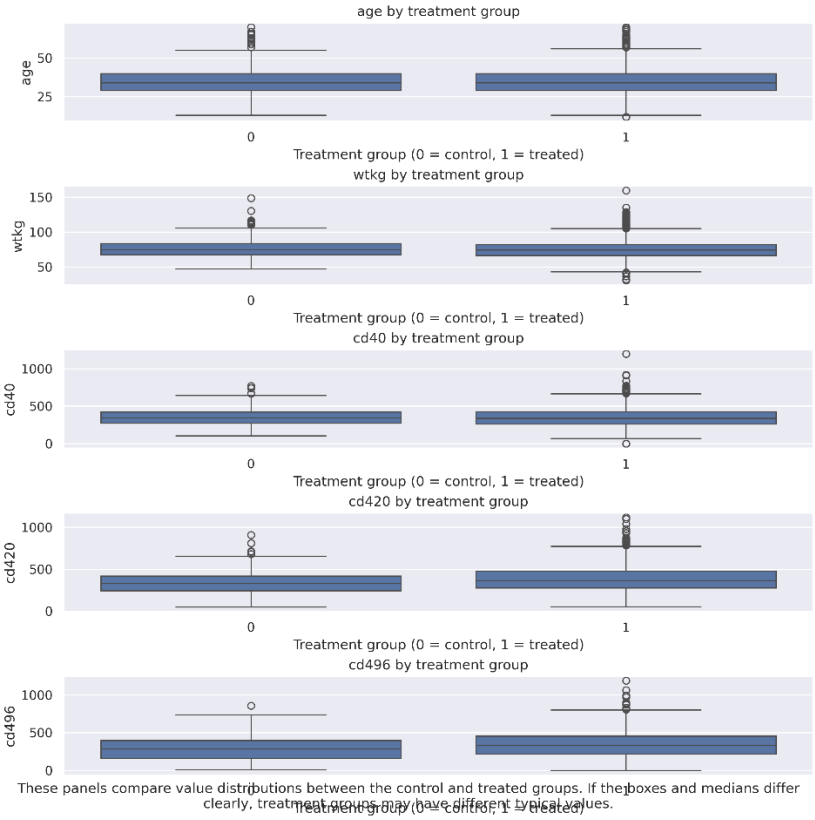
- **Figure S2-2:** Boxplots — continuous variables



Each box shows the middle half of the data (the 'typical' range). The line in the box is the median. Dots beyond the whiskers indicate unusually low or high values.

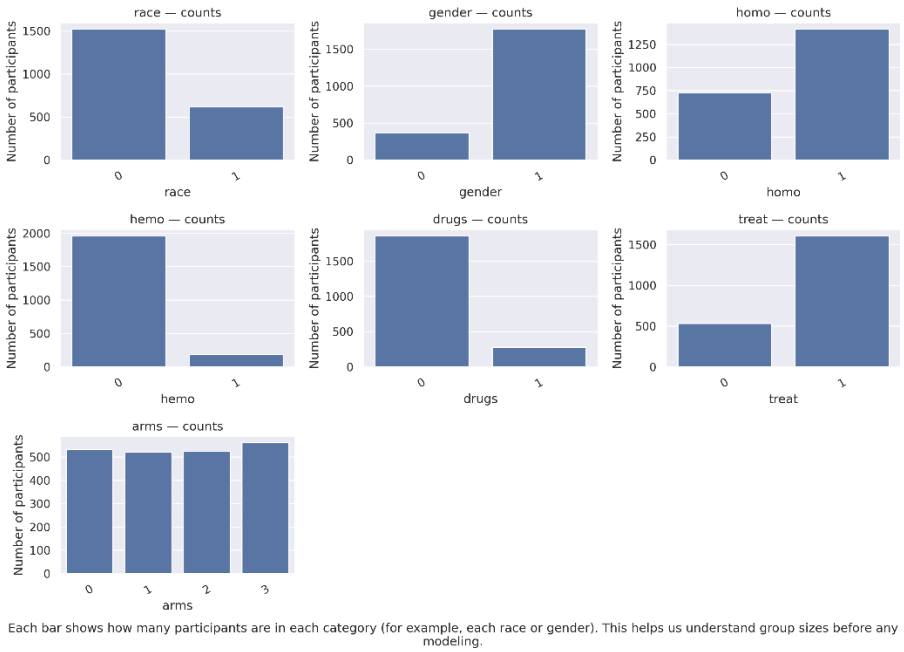
step1_and_step2/outputs/figs/02_boxplots_continuous.png

- **Figure S2-3:** Boxplots by treat



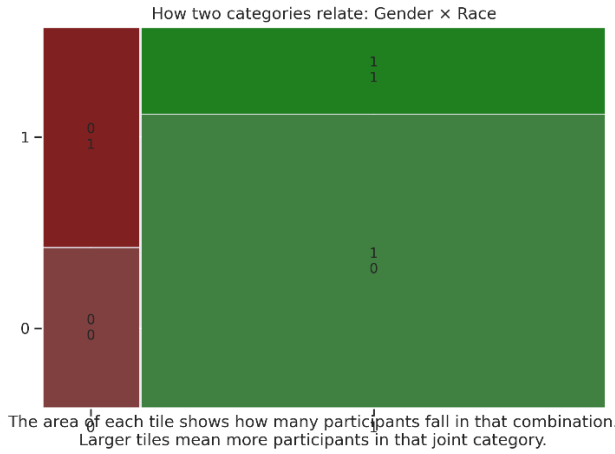
step1_and_step2/outputs/figs/03_boxplots_by_treat.png

- **Figure S2-4:** Bar charts — categorical variables



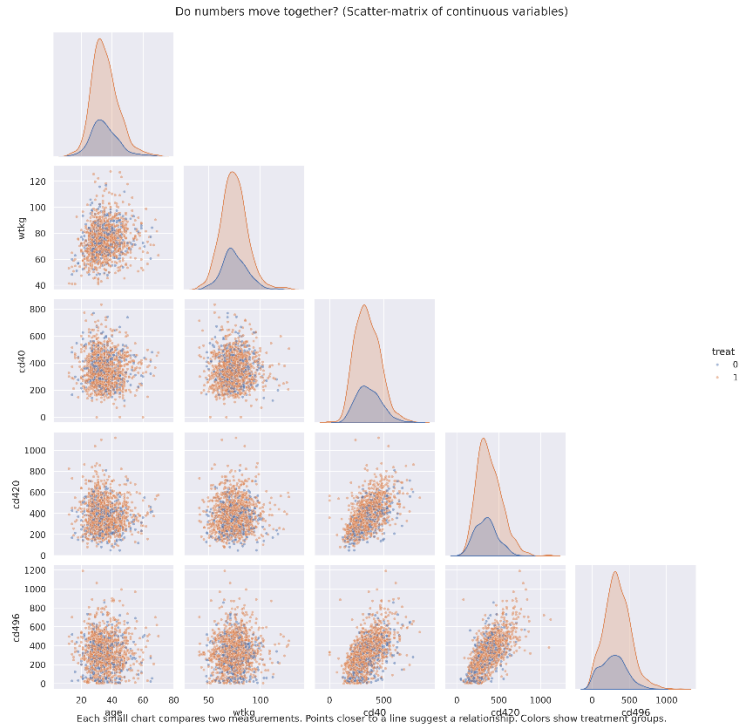
step1_and_step2/outputs/figs/04_barcharts_categorical.png

- **Figure S2-5:** Mosaic (Gender × Race)



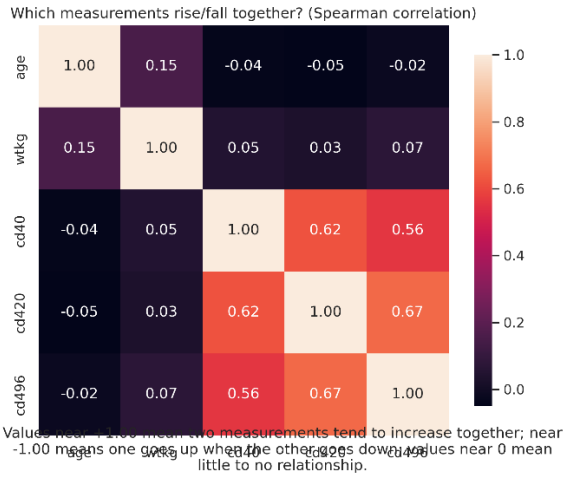
step1_and_step2/outputs/figs/05_mosaic_gender_by_race.png

- **Figure S2-6:** Pair plot — continuous subset



step1_and_step2/outputs/figs/06_pairplot_continuous.png

- **Figure S2-7:** Spearman correlation heatmap



step1_and_step2/outputs/figs/07_heatmap_spearman.png

Tables (from Step 1–2 CSVs)

- **Table S2-A:** Selected variables + type + sensitivity

	A	B	C
1	variable	role	sensitive
2	arms	categorical	no
3	drugs	categorical	yes
4	gender	categorical	yes
5	hemo	categorical	yes
6	homo	categorical	yes
7	race	categorical	yes
8	treat	categorical	no
9	age	continuous	no
10	cd40	continuous	yes
11	cd420	continuous	yes
12	cd496	continuous	yes
13	wtkg	continuous	no

step1_and_step2/outputs/csv/01_variables_selected.csv

- **Table S2-B:** Missingness (selected variables)

	A	B
1	variable	missing_count
2	age,	0
3	arms,	0
4	cd40,	0
5	cd420,	0
6	cd496,	797
7	drugs,	0
8	gender,	0
9	hemo,	0
10	homo,	0
11	race,	0
12	treat,	0
13	wtkg,	0

step1_and_step2/outputs/csv/02_missingness_subset.csv

- **Table S2-C:** Descriptive stats — continuous

	var	count	mean	std	min	25%	50%	75%	max
2	age	2139.0	35.24824684431977	8.70902623400872	12.0	29.0	34.0	40.0	70.0
3	wtkg	2139.0	75.12531051893409	13.263164003518359	31.0	66.6792	74.3904	82.5552	159.93936
4	cd40	2139.0	350.5011687704535	118.57386252156309	0.0	263.5	340.0	423.0	1199.0
5	cd420	2139.0	371.3071528751753	144.63490891153165	49.0	269.0	353.0	460.0	1119.0
6	cd496	1342.0	328.57078986587186	174.65615260925543	0.0	209.25	321.0	440.0	1190.0

step1_and_step2/outputs/csv/03_summary_continuous.csv

- **Table S2-D:** Counts — categorical

	variable	level	count
2	race	0	1522
3	race	1	617
4	gender	1	1771
5	gender	0	368
6	homo	1	1414
7	homo	0	725
8	hemo	0	1959
9	hemo	1	180
10	drugs	0	1858
11	drugs	1	281
12	treat	1	1607
13	treat	0	532
14	arms	3	561
15	arms	0	532
16	arms	2	524
17	arms	1	522

step1_and_step2/outputs/csv/04_counts_categorical.csv

- **Table S2-E:** Crosstab Gender×Race

	A	B	C
1	gender	0	1
2	0	155	213
3	1	1367	404

step1_and_step2/outputs/csv/05_crosstab_gender_race.csv

- **Table S2-F:** Spearman correlations

	A	B	C	D	E	F
1	variable	age	wtkg	cd40	cd420	cd496
2	age	1.0	0.151816646849143	-0.04003396359809793	-0.04986641458451257	-0.019500633453549094
3	wtkg	0.151816646849143	1.0	0.0474552672569389	0.03192548183872741	0.07066068466704756
4	cd40	-0.04003396359809793	0.0474552672569389	1.0	0.6200020385418676	0.5571459195284332
5	cd420	-0.04986641458451257	0.03192548183872741	0.6200020385418676	1.0	0.66901020152516
6	cd496	-0.019500633453549094	0.07066068466704756	0.5571459195284332	0.66901020152516	1.0

(CSV) step1_and_step2/outputs/csv/06_correlations_spearman.csv

Terminal output (how to read)

- Printed **column info**, **descriptive statistics**, **type inference**, and a concise list of **sensitive variables**.
- When it says "**SAVED:** .../figs/...png" or ".../csv/...csv", that's your confirmation the artifact was created.

3) Disclosure risk analysis (Step 3, R)

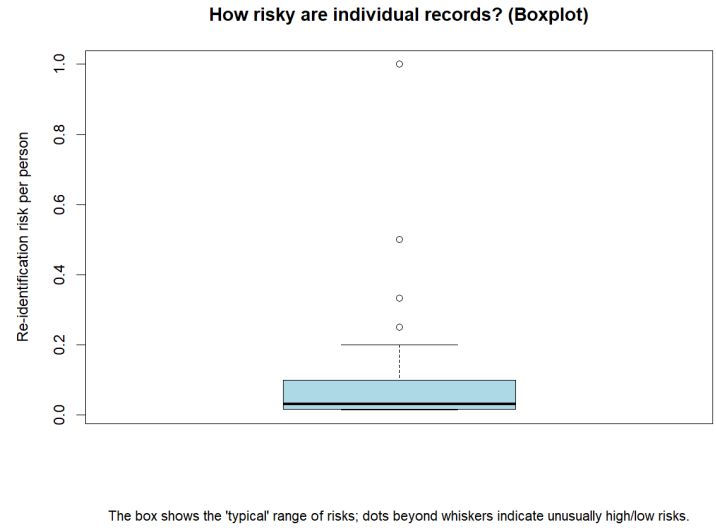
What we measured (R + sdcMicro)

Quasi-identifiers (QIs): **age, gender, race.**
Computed:

- **Global risk** (sdcMicro's global): **8.51%**
- **Expected re-identifications: 182 / 2139**
- **k-anonymity profile: % unique (k=1) = 1.36%, % with k ≤ 5 = 10.52%**
- **Inference risk:** arms → treat determinism

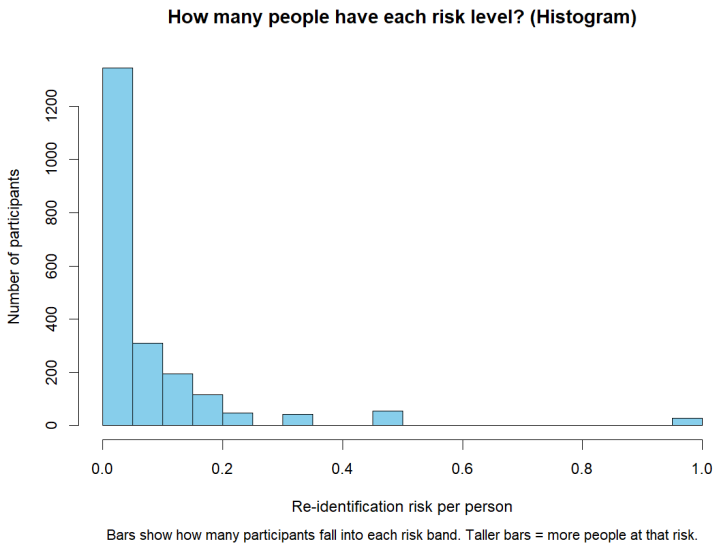
Figures (Step 3 PNGs)

- **Figure S3-1:** Boxplot — individual risk



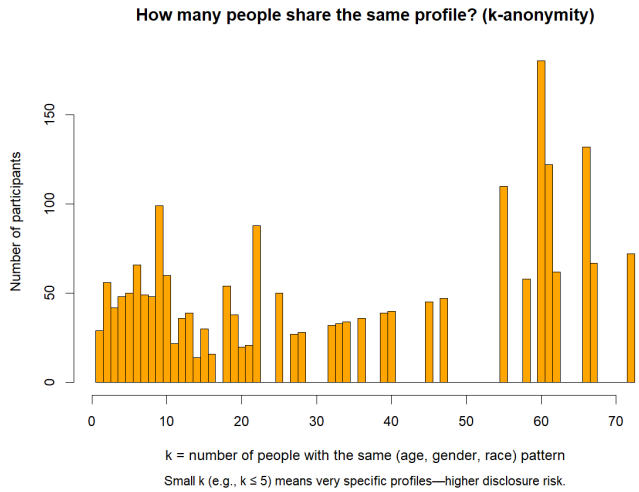
outputs/plots/01_boxplot_individual_risk.png

- **Figure S3-2:** Histogram — individual risk



outputs/plots/02_hist_individual_risk.png

- **Figure S3-3:** Histogram — equivalence class sizes (k)



outputs/plots/03_hist_equivalence_class_sizes_k.png

Tables (Step 3 CSVs)

- **Table S3-A:** Risk summary

	A	B	C	D	E
1	global_risk_percent	"expected_reidentifications"	"percent_unique_on_keys"	"percent_k_le_threshold"	"threshold_k"
2	8.51	182	1.36	10.52	5

outputs/csv/01_risk_summary.csv

- **Table S3-B:** Top-10 riskiest classes (by 1/k)

	A	B	C	D	E
1	age	"gender"	"race"	"k"	"individual_risk"
2	61	0	0	1	1
3	60	0	1	1	1
4	46	0	1	1	1
5	59	1	1	1	1
6	54	1	1	1	1
7	57	0	0	1	1
8	66	1	0	1	1
9	58	0	0	1	1
10	58	1	1	1	1
11	63	0	0	1	1
12					

outputs/csv/02_top10_risky_classes.csv

- **Table S3-C:** Small-count levels (per key var)

	A	B	C
1	variable	"level"	"count"
2	age	66	1
3	age	69	1
4	age	61	2
5	age	64	2
6	age	67	2
7	age	68	2
8	age	70	2
9	age	12	3
10	age	13	3
11	age	15	3
12	age	60	3
13	age	65	3
14	age	17	4
15	age	56	5
16	age	58	5
17	age	62	5

outputs/csv/03_small_count_levels_keyvars.csv

- **Table S3-D:** Risk snapshots across key sets

	keys	"pct_unique"	"median_k"	"pct_k_le_T"
2	age	0.09	90	2.15
3	gender	0	1771	0
4	race	0	1522	0
5	age+gender	0.56	56	5
6	age+race,	0.37	42	5.47
7	gender+race	0	1367	0
8	age+gender+race	1.36	33	10.52

outputs/csv/04_risk_snapshots_keysets.csv

- **Table S3-E:** Crosstab arms × treat (inference)
  (CSV) outputs/csv/05_crosstab_arms_treat.csv

Terminal output (how to read)

- Lines like Global risk: 8.51% and Expected re-identifications: 182 are headline risk.
- "Percent unique on (age, gender, race)" and "Percent with k ≤ 5" describe the **k-anonymity** profile.
- The 2-way table for arms × treat reveals the 1-to-1 mapping (attribute disclosure).
- Every CSV/PNG prints a **SAVED** message so you can copy the file paths directly into the report.

---

4) Anonymization experiments (Step 4, Python)

Methods & parameters

- **Age banding:** widths = 5, 10, 15 years
- **PRAM (binary flipping):** gender or race with flip prob p = 1%, 5%, 10%

Risk metrics (on QIs = age, gender, race)

- % unique, % with k ≤ 5, **expected re-IDs** (≈ sum of 1/k), **avg linkage risk** (mean 1/k × 100)

Utility metrics

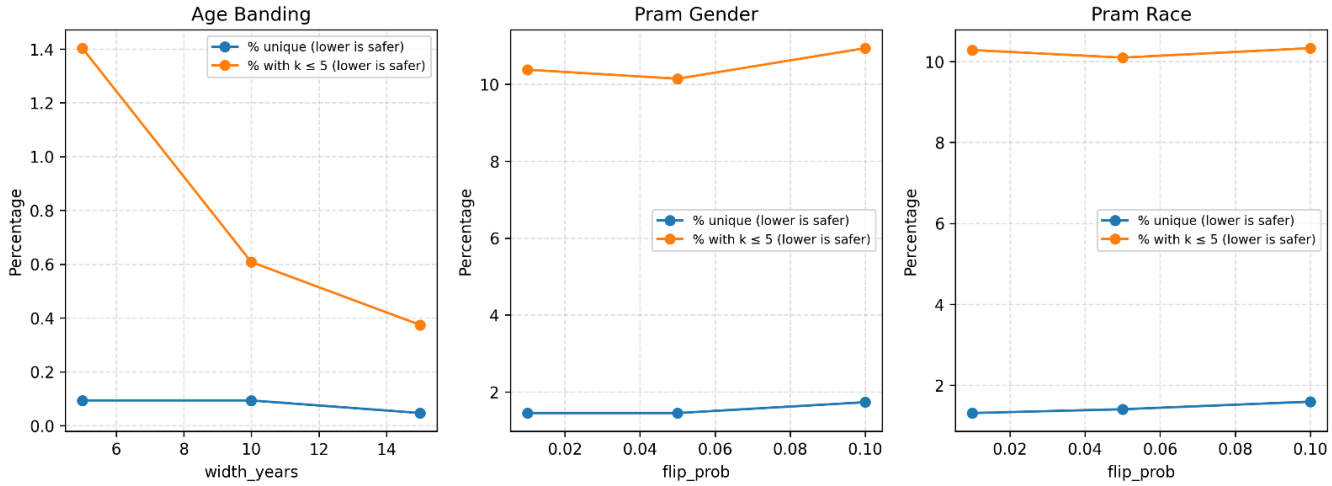
- **IL1 overall** (numeric component = mean normalized |Δ|; categorical = fraction flipped)
- **Eigenvalue similarity (%)** between correlation matrices

Key findings (from your run)

- **Age banding** markedly reduces risk with mild IL1 and minimal structure distortion:
  - **5-year:** expected re-IDs → **44**, k≤5 → **1.40%**, IL1 ≈ **0.021**, eigen-sim ≈ **99.94%**
  - **10-year:** expected re-IDs → **25**, k≤5 → **0.61%**, IL1 ≈ **0.044**, eigen-sim ≈ **99.75%**
  - **15-year:** expected re-IDs → **18**, k≤5 → **0.37%**, IL1 ≈ **0.067**, eigen-sim ≈ **99.39%**
- **PRAM** on a single binary key provides negligible risk reduction (age dominates), but adds categorical IL1.

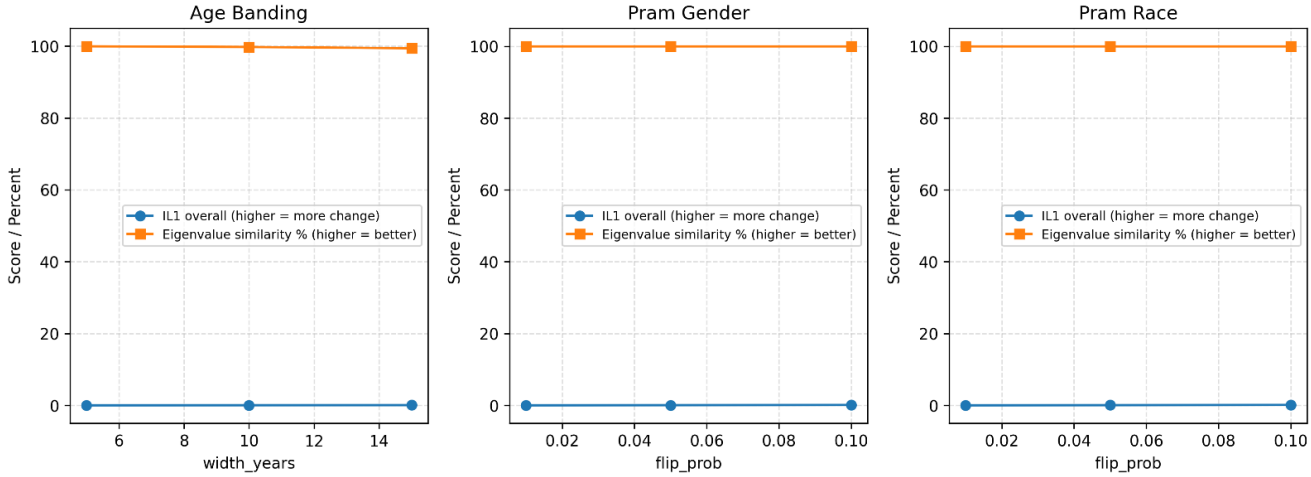
Figures (Step 4 PNGs in current folder)

- **Figure S4-1:** Risk vs parameter by method



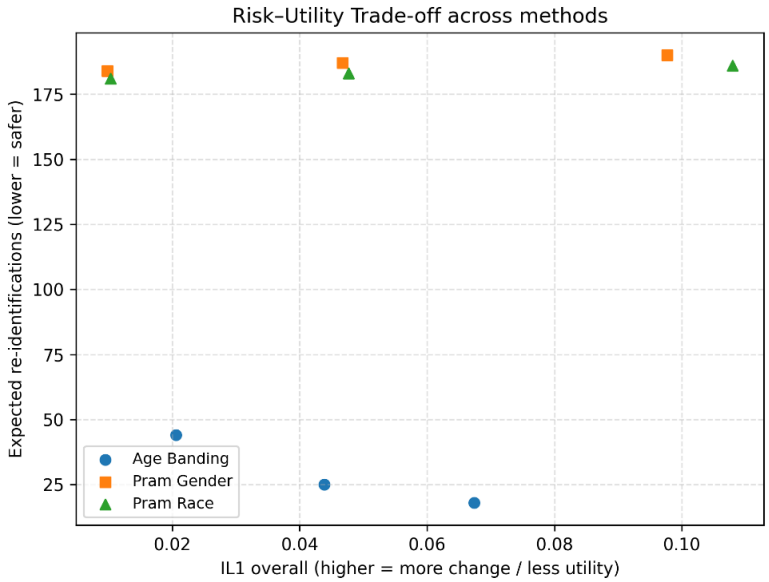
step4/step4_plot_risk_by_method.png

- **Figure S4-2:** Utility vs parameter by method



step4/step4_plot_utility_by_method.png

- **Figure S4-3:** Risk–Utility trade-off (all methods)



- step4/step4_plot_risk_utility_tradeoff.png

Tables (Step 4 CSV)

- **Table S4-A:** Summary of all runs (baseline + methods × params)

	A	B	C	D	E	F	G	H	I	J	K	L
1	method	param_name	param_value	percent_unique	percent_k_le_5	avg_linkage_risk_percen	expected_reids	n_keys_evaluated	IL1_numeric	IL1_categorical	IL1_overall	eigenvalue_similarity_percent
2	BASELINE			1.3557737260402059	10.518934081346423	8.508648901355773	182.0	2139	0.0	0.0	0.0	100.0
3	AGE_BANDING	width_years	5	0.09350163627863488	1.402524544179523	2.0570359981299675	44.0	2139	0.020570359981299677	0.0	0.020570359981299677	99.93706208855507
4	AGE_BANDING	width_years	10	0.09350163627863488	0.6077606358111267	1.168770453482936	25.0	2139	0.04385710370621141,	0.0	0.04385710370621141	99.74682945827294
5	AGE_BANDING	width_years	15	0.04675081813931744	0.3740065451145395	0.8415147265077139	18.0	2139	0.06742596443713626	0.0	0.06742596443713626	99.39224354378824
6	PRAM_GENDER	flip_prob	0.01	1.4492753623188406	10.37868162692847	8.60215053763441	184.0	2139	0.0	0.009817671809256662	0.009817671809256662	100.0
7	PRAM_GENDER	flip_prob	0.05	1.4492753623188406	10.144927536231885	8.742402992052362	187.0	2139	0.0	0.046750818131931744	0.046750818131931744	100.0
8	PRAM_GENDER	flip_prob	0.1	1.729780271154745	10.93969144460028	8.882655446470315	190.0	2139	0.0	0.09770920991117345	0.09770920991117345	100.0
9	PRAM_RACE	flip_prob	0.01	1.3090229079008884	10.285179990649837	8.461898083216457	181.0	2139	0.0	0.010285179990649837	0.010285179990649837	100.0
10	PRAM_RACE	flip_prob	0.05	1.402524544179523	10.098176718092567	8.55539971949509	183.0	2139	0.0	0.047685834502103785	0.047685834502103785	100.0
11	PRAM_RACE	flip_prob	0.1	1.5895278167367928	10.331930808789155	8.695652173913043	186.0	2139	0.0	0.10799438990182328	0.10799438990182328	100.0
12												

step4/step4_results_summary.csv

Terminal output (how to read)

- First prints baseline risk; then **SAVED** paths for the combined CSV and all PNGs.
- The printed head of step4_results_summary.csv lets you see exact numbers for each setting.

---

5) Risk metrics — definitions & math (deep dive)

Let Xbe the quasi-identifier vector (here: age, gender, race).
For each record i, define its **equivalence class**:

$$\mathcal{C}(i) \;=\; \{j\colon X_j = X_i\}$$

with **class size** $k_i = \mid \mathcal{C}(i) \mid$.

k-Anonymity profile

- **% unique:** $\frac{1}{n}\sum_{i=1}^{n} \quad \mathbf{1}[k_i = 1] \times 100\%.$
- **% with k ≤ 5:** $\frac{1}{n}\sum_{i=1}^{n} \quad \mathbf{1}[k_i \le 5] \times 100\%.$

Linkage risk & Expected re-IDs
Assume an attacker who only knows QIs and guesses uniformly among matches.

- **Per-record success probability:** $p_i = 1/k_i.$
- **Average linkage risk (%):** $\frac{100}{n}\sum_i \quad p_i.$
- **Expected re-identifications:** $\sum_i \quad p_i.$
  These are exactly what you reported in Step 3 & 4 (and what sdcMicro's global/ER metrics embody when using QIs).

Attribute disclosure
If a sensitive attribute S(here treat) is a deterministic function of a released variable Z(here arms), i.e., $S = f(Z)$, then **attribute disclosure risk = 100%** for anyone in the class (knowing Zreveals S). Your contingency table confirmed $P(S = 1 \mid Z) \in \{0,1\}$.

6) Utility / information-loss metrics — definitions & math

IL1 (overall)
We used a transparent, interpretable IL1:

- **Numeric IL1:** For each numeric variable c:

$$\text{IL1}_c^{num} = \frac{1}{m_c} \sum_{i \in \text{nonNA}} \frac{| x'_{ic} - x_{ic} |}{\max(x_c) - \min(x_c)}$$

  where $m_c$is the number of non-missing aligned pairs; the denominator normalizes to the original range.
  The **numeric IL1** is the mean across all changed numeric variables.

- **Categorical IL1:** For each categorical c:

$$\text{IL1}_c^{cat} = \frac{1}{m_c} \sum_{i \in \text{nonNA}} \mathbf{1}[x'_{ic} \neq x_{ic}]$$

  The categorical IL1 is the mean across changed categoricals.

- **Overall IL1:** mean of the (available) numeric & categorical components.
  Interpretation: **0** = no change; higher = more distortion (less utility).

Eigenvalue similarity (%) of correlation matrices
Let $R$and R'be the correlation matrices on a shared set of numeric variables (pairwise deletion if needed); let $\lambda$and λ'be their eigenvalue vectors sorted descending. We use an L1 similarity:

$$\text{EigenSim}(\%) \;=\; 100\left(1 - \frac{\| \lambda' - \lambda \|_1}{\sum_j \quad \lambda_j}\right)$$

For a $p \times p$correlation matrix, $\sum_j \quad \lambda_j = p$. Values near **100%** mean the **multivariate structure** is preserved.

7) Interpreting the privacy–utility trade-off

- **Age banding** improves anonymity by pooling individuals into larger equivalence classes (bigger k), lowering **% unique**, **% with k ≤ 5**, and **expected re-IDs**.
- IL1 rises gradually with wider bands because age values move further from their originals; however, at **10-year bands**, the **utility loss is still small** (≈0.044) and **correlation structure remains ~99.75% intact**.

- **PRAM on one binary QI** barely changes the equivalence classes; risk barely moves while categorical IL1 increases, so the trade-off is unfavorable here.
- **Bottom line: 10-year age bands** give an excellent **risk–utility compromise**; if you need stronger privacy, **15-year bands** are still quite gentle on utility.
- **Critical:** remove or coarsen variables that **deterministically reveal** a sensitive attribute (here, avoid releasing both arms and treat).

---

8) Code snippets (for documentation & reproducibility)

8.1 — R (Step 3 essentials)

```
# Risk on QIs with sdcMicro
library(sdcMicro)
df <- read.delim("aids_original_data.csv", sep=";")
key_vars <- c("age","gender","race")
sdc <- createSdcObj(dat=df, keyVars=key_vars, sensibleVar="treat")

rk  <- get.sdcMicroObj(sdc, type="risk")
global_risk   <- rk$global$risk
expected_reid <- rk$global$risk_ER

# k-anonymity sizes
fk <- freqCalc(df, keyVars=key_vars)$fk
pct_unique <- mean(fk==1)*100
pct_k_le_5 <- mean(fk<=5)*100
```

8.2 — Python (Step 4 essentials)

```
import numpy as np, pandas as pd

def k_equivalence_sizes(keys_df):
  counts = keys_df.value_counts(dropna=False)
  return keys_df.apply(lambda row: counts[tuple(row)], axis=1)

def risk_metrics(df, qis):
  sub = df[qis].dropna()
  k = k_equivalence_sizes(sub)
  return {
    "percent_unique": (k.eq(1).mean()*100),
    "percent_k_le_5": (k.le(5).mean()*100),
    "expected_reids": float((1.0/k).sum()),
    "avg_linkage_risk_percent": (1.0/k).mean()*100
  }

def age_band(s, width):
  return (np.floor(s/width)*width + width/2).astype(int)

# Example: 10-year banding
df = pd.read_csv("aids_original_data.csv", sep=";").rename(columns=str.lower)
df_10 = df.copy()
df_10["age"] = age_band(df_10["age"], 10)

print(risk_metrics(df, ["age","gender","race"]))
print(risk_metrics(df_10, ["age","gender","race"]))
```

8.3 — Reading your Step 4 master table

```
import pandas as pd
s = pd.read_csv("step4_results_summary.csv")
# Compare methods at a glance
print(s.pivot_table(index=["method","param_value"],
                    values=["percent_unique","percent_k_le_5",
                            "expected_reids","IL1_overall",
                            "eigenvalue_similarity_percent"]))
```

---

9) Final recommendations

1. **Publish with 10-year age bands**; consider 15-year bands if your risk tolerance requires.
2. **Do not release both arms and treat.** If treat must be released, coarsen or remove arms.
3. Keep your **risk table & trade-off figures** in the appendix so readers can see exactly how parameter choices move privacy and utility.