Working with `sdcMicro`

From risky to shareable — anonymizing a real dataset, live

Prof. Dr. Matthias Templ

FHNW ▪ SwissAnon ▪ Lead developer of `sdcMicro`

Data Protection Day 2026 ▪ University of Zurich

Removing names is where it starts, not where it ends

- You stripped names, e-mails, birthdates. Done? No — that is where the real work begins.
- **Pseudonymization**: replace identifiers with artificial ones ("Hans Bauer" → "Participant1"). Reversible with the key.
- **Anonymization**: irreversible — data subjects can no longer be identified.
- Even with direct identifiers gone, quasi-identifiers (age, sex, location, . . .) can re-identify records.
- Sweeney (2000): 87 % of Americans are unique on ZIP code, date of birth and sex.
- Under the GDPR, pseudonymized data is still personal data.

Every variable plays one role

- **Direct identifiers:** identify a person directly — names, passport numbers, addresses. *(Remove.)*
- **Quasi-identifiers:** age, sex, occupation, location, . . . — do not identify alone, but re-identify in combination or via linkage to external data. *(Protect.)*
- **Confidential / sensitive variables:** health status, substance use, attitudes — what we must keep from leaking.
- **Non-confidential variables:** colour preference, reaction time — harmless on their own.

Measuring disclosure risk

k-Anonymity (Sweeney, 2002)

- Each combination of quasi-identifiers appears in at least k records.
- Nobody should be unique on the released quasi-identifiers ($k \geq 2$).
- Protects against identity disclosure (finding *who*).

ℓ-Diversity (Machanavajjhala et al., 2007)

- Each k-anonymous group must also be diverse in the sensitive variable.
- A group can satisfy k-anonymity yet reveal the same outcome for everyone in it.
- Protects against attribute disclosure (learning *what*).

Example: k-anonymity and ℓ-diversity

Fictitious infant dataset: a unique record ($k = 1$), a protected group ($k = 2$, $\ell = 2$), and a group with attribute disclosure ($k = 2$, $\ell = 1$).

Quasi-identifiers			Sensitive variable	k	ℓ
Age (months)	**Race**	**Gender**	**Newborn health condition**		
0–3	Asian	Female	Low birth weight	1	1
0–3	Black	Female	Colic	2	2
0–3	Black	Female	Low birth weight	2	2
4–6	White	Male	Lactose intolerance	2	1
4–6	White	Male	Lactose intolerance	2	1

The `sdcMicro` toolbox — three families we will use live

Today, on the Coleman data:

- Recoding / generalization — age bands, merge rare categories
- Suppression — replace the few remaining risky values with `NA` to reach k-anonymity
- Perturbation — PRAM: swap categories under an invariant transition matrix

Also in the box (mentioned, not shown):

- Rank swapping & added noise
- Synthetic data
- Record swapping, microaggregation, ...

Everything via the `sdcMicro` package (Templ et al., 2015) or its GUI `sdcApp` (Meindl & Templ, 2019).

Our case: the Coleman depression study

The data

- $n = 107$ participants, UK sample — real psychology data
- 4 quasi-identifiers: `Age`, `Sex`, `Occupation`, `Ethnicity`
- 47 sensitive Likert items → depression (PHQ-9), derailment, self-criticism, self-reassurance
- Direct identifiers already removed

Why this dataset

- The everyday case: small-n human-subjects data
- Small enough to read row by row
- Lets us watch k-anonymity fail — and fix it

Colman et al. (2024), PsychArchives; re-analysed in Müller, Templ & Strobl (2026).

Step 1 — write the disclosure scenario *first*

Our scenario (fixed before we touch a method)

Release mode **Public Use File** — openly downloadable, no data-use agreement.

Attacker knows `Age`, `Sex`, `Occupation`, `Ethnicity` of the target, and that they took part.

Harm identity disclosure ⇒ the attacker reads the participant's depression score.

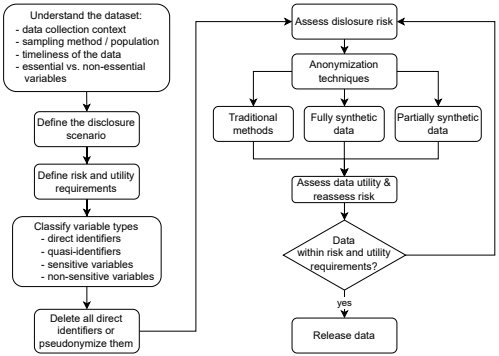
Utility the original regression must stay reproducible.

Without a written scenario you cannot decide *when to stop* — suppress one more cell? bands of 5 or 20 years?

56 % of these 107 people are unique on just those four variables.

Let us watch `sdcMicro` fix that — live.

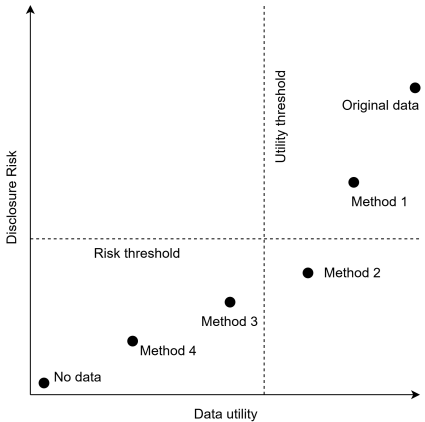
The workflow we run now — live in `sdcApp`



Risk → recode → suppress → perturb → check utility → document. *(Switch to `sdcApp()` now.)*

Risk vs. utility — there is no free lunch

- Every method trades information for protection.
- The goal is not zero risk, but an *acceptable balance*.
- The balance depends on the use case and the sensitivity of the data.
- Your written scenario is what makes "acceptable" concrete.



Take-home

- Removing names is not anonymization — quasi-identifiers remain.
- Write the disclosure scenario first; it tells you when to stop.
- Recode → suppress → perturb, then re-measure and document.
- `sdcMicro` gives you a *toolbox* and an *audit trail* — the report is the deliverable, not optional metadata.

Try it yourself

- `install.packages("sdcMicro")`
- `library(sdcMicro); sdcApp()` — launches the GUI
- Reports ship *with* the data: `report(sdc, internal = TRUE/FALSE)`

Go deeper: `sdcMicro` (Templ et al., 2015) ▪ `sdcApp` GUI (Meindl & Templ, 2019) ▪ Workshop 5 (data-sharing Q&A) ▪ find me afterwards — FHNW / SwissAnon.

Slides, script & data

swissanon.com/dpd2026

References

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), 1–52. https://doi.org/10.1145/1217299.1217302

Meindl, B., & Templ, M. (2019). Feedback-based integration of the whole process of data anonymization in a graphical interface. *Algorithms*, *12*(9), 191. https://doi.org/10.3390/a12090191

Sweeney, L. (2000). *Simple demographics often identify people uniquely* (Data Privacy Working Paper No. 3). Carnegie Mellon University. Pittsburgh, PA. Retrieved May 18, 2026, from https://privacytools.seas.harvard.edu/files/privacytools/files/paper1.pdf

References

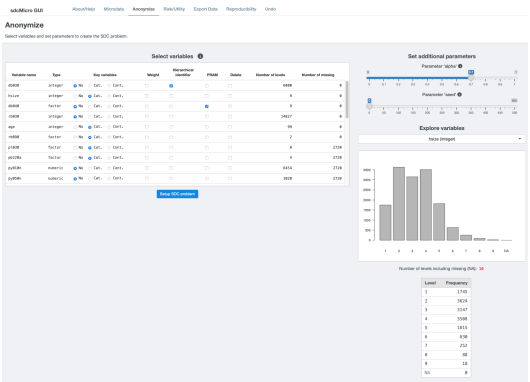
Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software*, *67*(4), 1–36. https://doi.org/10.18637/jss.v067.i04

Thank you

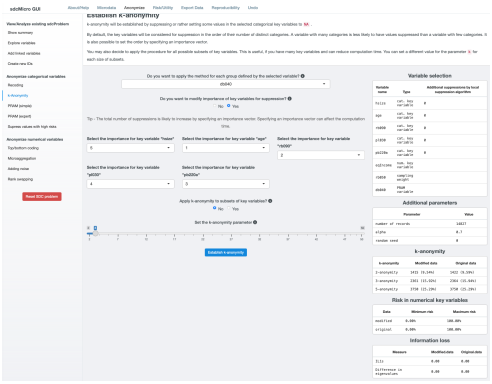
Questions?

Backup — setup: assign the key variables



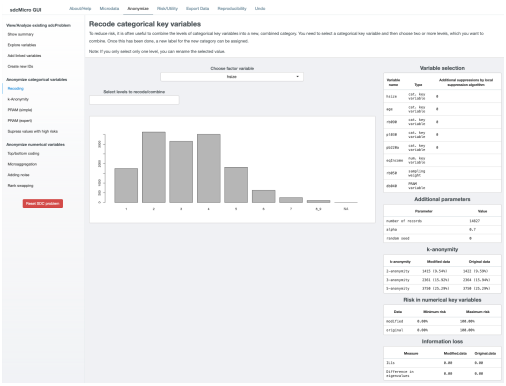
Anonymize data → select `Age`, `Sex`, `Occupation`, `Ethnicity` as key variables.

Backup — initial risk: 56 % sample-unique



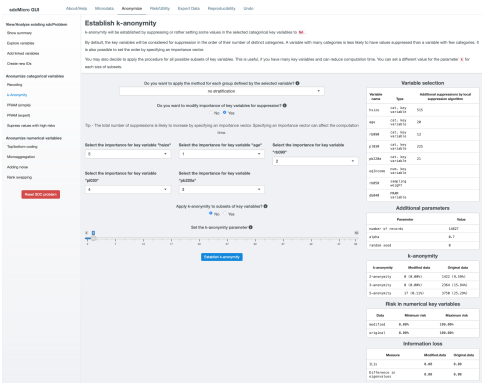
Risk → *k-Anonymity*. Red rows violate the chosen k.

Backup — recoding: age bands & rare categories



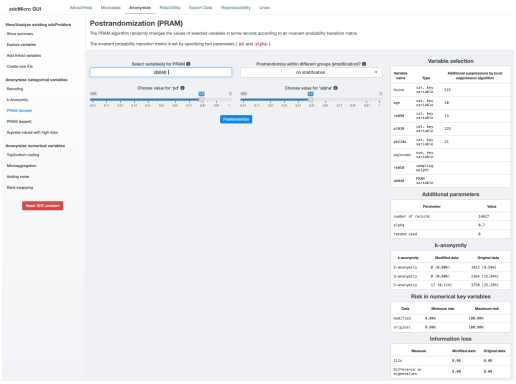
Anonymize → *Recoding* → global recoding / group categories.

Backup — local suppression to $k = 3$



Anonymize → *Local suppression*: set k and the importance vector, *Apply*.

Backup — PRAM on the categorical quasi-identifiers



Anonymize → *PRAM*: pick the variables, accept or edit the matrix, *Apply*.