data-obfuscator quick reference

Mask a NONMEM-style PK/PD dataset before showing it to a large language model or a collaborator. Everything can be given on the command line or in an options file; command-line flags override the file. Complete list with every default: CONFIGS.md.

The four commands

Command	Purpose
`run`	Mask one dataset. Writes the masked file, llm.txt, tokens.txt, the private key and a record of the run.
`run2`	Mask several related datasets together (for example CDISC ADaM ADSL, ADEX, ADPC, ADLB) with shared codes.
`reverse`	Undo the masking with the private key: restores a masked dataset, or the combined dataset an LLM built from run2 files.
`tokens`	Estimate how many tokens a file would use in the main LLM families, before you paste it.

Typical workflow

```
data-obfuscator run --input pk_study.csv --output ./masked --level 2
```

Check the warnings and masked/summary.txt. Share only obfuscated.csv and llm.txt. Keep keymap.txt, summary.txt, config.used.txt and obfuscator.log private. Restore with:

```
data-obfuscator reverse --input table_from_llm.csv --keymap masked/keymap.txt
```

Masking levels (each includes the one before)

Level	Name	What it masks
1	light	Identifiers: ID, subject (USUBJID), study, site and user columns are recoded.
2	clinical (default)	Level 1, plus treatment, sex, race and country labels, dates shifted by a constant, and free text redacted.
3	covariate	Level 2, plus covariates (WT, AGE, HT, CRCL, ...) multiplied or divided by a hidden random factor per column.
4	maximum	Level 3, but every value gets its own factor (the column factor plus or minus 15 %), free-text columns are dropped, and subjects are re-ordered.

Never altered at any level: TIME, TAD, DV, AMT, RATE, EVID, MDV, CMT, ADDL, II, SS, DVID, BLQ and similar columns.

run: masking one dataset

Option	Default	What it does
`--input FILE`	required	The dataset (CSV, TSV, semicolon or whitespace delimited).
`--output DIR`	required	Output folder. Must be new or empty unless --overwrite.
`--config FILE`	none	Options file.
`--level 1-4`	2	How much to mask (identifiers, + labels/dates/text, + covariates, maximum).
`--seed N`	42	Random seed. Same input + options + seed = identical output. Use your own seed when privacy matters.
`--id-style`	sequential	Style of recoded IDs: sequential, padded or hashed.
`--date-shift-days`	1-100	A window to draw the shift from (1-100), or one fixed number of days (23.6).
`--scale-range MIN-MAX`	1-20	Range of the random factor for scaled columns (levels 3 and 4).
`--scale-columns`	none	Extra numeric columns to scale. Covariates such as WT and AGE are scaled automatically.
`--keep / --drop`	all / none	Columns to keep, or to drop.
`--protect`	TIME, DV, AMT, EVID, ...	Columns that must never be altered.

Option	Default	What it does
`--freetext-columns, --drop-freetext`	auto	Free-text columns to redact, or drop instead of redacting.
`--rename-columns, --rename-except`	none	Rename listed columns (or every column except the listed ones).
`--write-keymap BOOL`	true	Write the private key. Without it the masking can never be undone.
`--overwrite`	off	Allow a non-empty output folder.

reverse: undoing the masking

Option	What it does
`--input FILE`	The masked dataset (or the master dataset built from run2 files). Rows and columns may be in any order.
`--keymap FILE`	keymap.txt or keymap2.txt (default: next to the input). The program recognises which it is.
`--keymap-columns FILE`	Level 4 only: keymap_columns.txt (default: next to keymap.txt).
`--output DIR`	Where to write restored.*, reverse_report.txt and reverse.log (default ./restored).
`--only / --skip`	Undo only, or all but, these parts: rename, scaling, dates, recoding.
`--strict`	Stop with an error instead of approximating or leaving unknown values unchanged.

Global options

--version (banner and a description of the program), --help (also available for every command), --no-color.

Masking lowers the risk of re-identification; it is not formal anonymisation. Read every warning the run prints and decide with your organisation's rules before sharing a dataset.