Quantifying Re-Identification Risk before Sharing Clinical Data: an Open Implementation and an Empirical Risk–Utility Evaluation

Dr. Shwetha Javali

Associate Professor, Department of Obg Nursing Bldea's Shri B M Patil Institute of Nursing Sciences Vijayapur, India.

Corresponding author: hitnalshwetha@gmail.com

Received: 05 June 2026; Revised: 13 August 2026; Accepted: 20 August 2026; Published: 22 September 2026

Abstract

Background: Clinical datasets are routinely shared after removing direct identifiers, on the assumption that what remains is anonymous. Whether it is depends on the combinatorics of the remaining variables, which is computable in advance and is rarely computed. Data custodians need a way to quantify the residual risk, to see what de-identification costs in analytic utility, and to choose defensibly between the two.

Objective: To describe a compact open implementation for measuring re-identification risk and generalisation-based de-identification of tabular clinical data, and to report an empirical evaluation of the resulting risk–utility trade-off.

Methods: The toolkit computes equivalence classes over a declared quasi-identifier set and reports k-anonymity, the proportion of unique records, prosecutor and mean record risk, l-diversity and the discernibility metric. It applies generalisation hierarchies to age, admission date and geography at four levels each, and record suppression to a target k. Utility is measured as information loss and as the cross-validated discrimination of a downstream outcome model. The toolkit was evaluated on a synthetic cohort of 50,000 admissions with six attributes, constructed so that the combinatorial structure resembles a routine hospital extract.

Results: Age alone gave k = 94 and no unique records. Adding sex reduced k to 43. Adding district produced 14.6 % unique records, and adding the exact admission date took uniqueness to 99.6 % — a single variable moving the dataset from partially protected to almost entirely identifiable. Generalisation to 10-year age bands, month of admission and 22 geographic units reduced uniqueness to 15.7 %; a further step to 20-year bands, year of admission and full geographic suppression eliminated unique records entirely, at an information loss of 31 % and with downstream discrimination unchanged (AUROC 0.586 to 0.593). Reaching k = 10 by suppression alone at the intermediate generalisation level required discarding 86.7 % of records, against 0.08 % at the coarser level. Uniqueness remained above 99 % at every cohort size from 1,000 to 200,000.

Conclusions: Risk is driven by a small number of high-cardinality variables, notably exact dates, and is correctable by generalising those variables at negligible cost to a typical downstream analysis. Suppression is an expensive substitute for generalisation. Larger cohorts confer no protection.

Keywords: Data Privacy, Re-Identification Risk, K-Anonymity, De-Identification, Health Data Sharing, Open-Source Tools.

1. BACKGROUND

Sharing clinical data for secondary research is a governance problem before it is a technical one, and the governance usually turns on a single word. A dataset with names, identifiers and contact details removed is described as de-identified or anonymised, and the two terms are treated as interchangeable. They are not. Removing direct identifiers eliminates the easy route to a patient; it says nothing about whether the combination of variables that remains — age, sex, area of residence, dates of contact, diagnosis — picks out an individual uniquely.

That question is combinatorial and therefore computable. Records sharing the same values on a declared set of quasi-identifiers form an equivalence class; a record in a class of one is unique and, if an adversary knows those attributes about a target, re-identifiable with certainty. The size of the smallest class in a dataset is its k-anonymity, and the distribution of class sizes determines the residual risk more informatively than any single number.

These ideas are three decades old and well represented in the privacy literature, and the operational gap is not conceptual. It is that a data custodian deciding whether to release an extract next week needs a number, needs to know what protecting the data will cost the people who want to analyse it, and often has neither a tool nor a quantified basis for the trade-off. Institutional practice consequently defaults to fixed rules — remove identifiers, band the ages, keep the month — applied uniformly regardless of what the data actually look like.

This report describes a small implementation intended to close that gap, and reports what it found on a worked evaluation. Three questions are addressed. Which variables actually drive risk in a typical extract? What does generalisation cost the analyst? And how do generalisation and suppression compare as routes to a given level of protection?

1.1 Threat Models

Risk is meaningless without an adversary, and the three conventionally distinguished are computed separately by the toolkit because they answer different governance questions.

Table 1. Threat models and the corresponding risk measure

Threat model	Adversary	Question answered	Measure
Prosecutor	Knows a specific individual is in the dataset and knows their quasi-identifiers	What is the worst case for any one person?	1 / k, where k is the smallest equivalence class
Journalist	Wants to re-identify someone, anyone, to demonstrate that the release was unsafe	How exposed is the most vulnerable record?	Proportion of unique records
Marketer	Wants to re-identify as many records as possible; does not care about any individual	What is the expected yield across the dataset?	Mean of 1 / class size across records

The prosecutor model is the strictest and the one regulators generally have in mind; the marketer model produces the smallest numbers and is the one most likely to be quoted by a party that wishes to release. Reporting all three makes the choice of model explicit rather than implicit.

2. IMPLEMENTATION

2.1 Design

The toolkit is deliberately small: a single module, no dependencies beyond the standard scientific Python stack, and no database or service component. It is intended to be read and audited by a data custodian rather than trusted as a black box, and to be run on an extract immediately before release.

Table 2. Components

Component	Function	Output
Equivalence class computation	Groups records by the declared quasi-identifier set	Class sizes; class count
Risk metrics	Computes k, uniqueness, prosecutor, journalist and marketer risk	Risk profile for the release
l-diversity	Smallest number of distinct sensitive values within any class	Attribute-disclosure indicator

Discernibility metric	Sum of squared class sizes, normalised	Scalar utility penalty for coarse grouping
Generalisation hierarchies	Four levels each for age, date and geography	Transformed dataset
Suppression	Removes records in classes below a target k	Filtered dataset; percentage suppressed
Utility evaluation	Information loss; cross-validated AUROC of a downstream model	Analytic cost of the transformation

2.2 Risk Measures

For a dataset of n records partitioned into equivalence classes of sizes $s_1 \dots s_m$, the toolkit computes

$k = \min s_j$, prosecutor risk = $1/k$, mean record risk = $(1/n)\ \Sigma\ s_j \times (1/s_j) = m/n$ (1)

together with the proportion of records in classes of size one and the proportion in classes smaller than five, the latter being the threshold used in many national data-release policies. The discernibility metric, the normalised sum of squared class sizes, rises as generalisation coarsens and provides a utility penalty that does not require a downstream model.

2.3 Generalisation Hierarchies

Table 3. Generalisation levels applied in the evaluation

Level	Age	Admission date	Geography
L0	Exact year	Exact day	220 districts
L1	5-year bands	Week	44 units of 5 districts
L2	10-year bands	Month	10 units of 22 districts
L3	20-year bands	Year	Suppressed entirely

The hierarchies are conventional and were fixed before any risk was computed. Levels may be combined independently, which gives the 64 combinations explored in Section 3.4; the evaluation in Sections 3.2 and 3.3 applies the same level to all three variables for clarity.

2.4 Evaluation Cohort

The evaluation cohort is synthetic and is described as such throughout: It comprises 50,000 admissions with age, sex, district of residence drawn from a skewed 220-district distribution, admission day within a three-year window, ethnicity in five categories, specialty in six, and a binary outcome generated from age, sex and specialty with substantial noise. No real patient data were used.

Synthetic data are adequate for the claims made here and inadequate for others, and the distinction matters. Re-identification risk is a property of the combinatorial structure of a table — the cardinality of each variable and the sparsity of the joint distribution — not of the clinical truth of its contents, so the risk results transfer to any real extract with comparable structure. Claims about the specific risk of any real dataset would not transfer, and none is made. The downstream utility results are weaker on this point and are qualified in Section 4.3.

3. EVALUATION

3.1 Which Variables Drive Risk

Table 4 and Figure 1 add quasi-identifiers one at a time to an otherwise unmodified dataset.

Table 4. Risk as quasi-identifiers accumulate (n = 50,000, no generalisation)

Quasi-identifier set	k	Classes	Unique (%)	In classes < 5 (%)	Mean record risk
Age	94	81	0.00	0.00	0.0016
+ sex	43	162	0.00	0.00	0.0032

+ district	1	17,978	14.61	55.75	0.3596
+ admission day	1	49,902	99.61	100.00	0.9980
+ ethnicity	1	49,963	99.85	100.00	0.9993

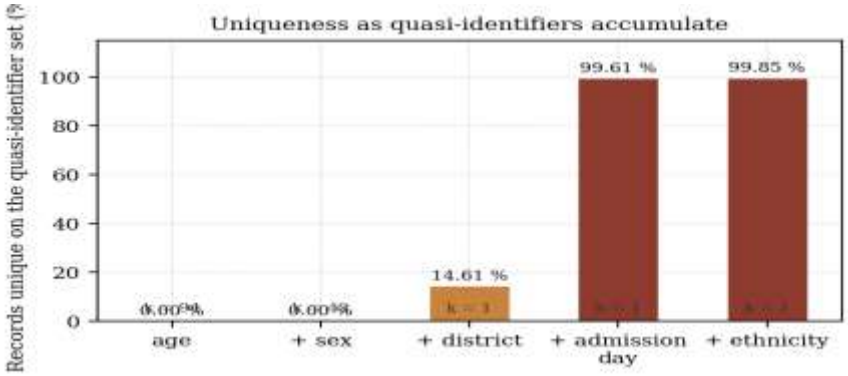

Figure 1. Proportion of records unique on the quasi-identifier set as variables accumulate, with the corresponding k.

The progression is sharply non-linear and the fourth row is the finding. Age and sex together leave every record in a class of at least 43 and nothing unique. Adding district of residence — 220 categories with a skewed distribution — makes 14.6 % of records unique and puts more than half in classes below five. Adding the exact admission date takes uniqueness to 99.6 %.

A single high-cardinality variable therefore moved the extract from partially protected to effectively fully identifiable, and it is the variable most likely to survive a conventional de-identification pass, because a date is not an identifier and analysts ask for it. The final row shows how little is left to lose: adding ethnicity, a five-category variable that governance review would scrutinise closely, raises uniqueness by a further 0.24 percentage points. Attention in de-identification review is routinely allocated in inverse proportion to where the risk lies.

3.2 What Generalisation Achieves, and What It Costs

Table 5. Risk and utility across generalisation levels (all three variables at the same level)

Level	k	Classes	Unique (%)	In classes < 5 (%)	Mean risk	Info. Loss (%)	AUROC
L0 exact	1	49,963	99.85	100.00	0.9993	0.0	0.586
L1	1	47,010	88.45	99.99	0.9402	1.5	0.585
L2	1	18,373	15.71	55.26	0.3675	4.7	0.580
L3	2	150	0.00	0.02	0.0030	31.3	0.593

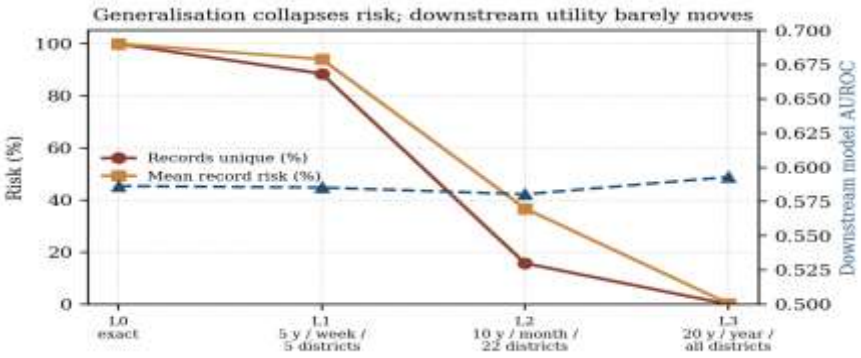

Figure 2. Risk measures (left axis) and downstream model discrimination (right axis) across generalisation levels.

Risk falls by three orders of magnitude between L0 and L3 while downstream discrimination does not fall at all: AUROC moves from 0.586 to 0.593, and the small rise is within the noise of the cross-validation. The information-loss figure of 31 % at L3 measures the coarsening of the cells, not the destruction of analytic value, and the two are not the same thing.

The reason is worth stating because it generalises beyond this cohort. The outcome in these data depends on age, sex and specialty. Age survives 20-year banding with most of its signal intact, because the relationship is monotone and smooth; sex and specialty are untouched by the hierarchies. The variables whose generalisation destroys risk — exact date and fine geography — carry almost no signal for this outcome. Where the risk-bearing variables and the signal-bearing variables are different variables, the trade-off is not a trade-off at all, and a custodian who assumes one exists will over-protect analysts against a cost they are not paying.

The converse case is real and the toolkit will detect it: an analysis of seasonality, of geographic clustering or of time-to-event on a daily scale depends on exactly the variables that must be coarsened, and for those the trade-off binds hard. The point is that which case applies is an empirical question about the intended analysis, answerable before release.

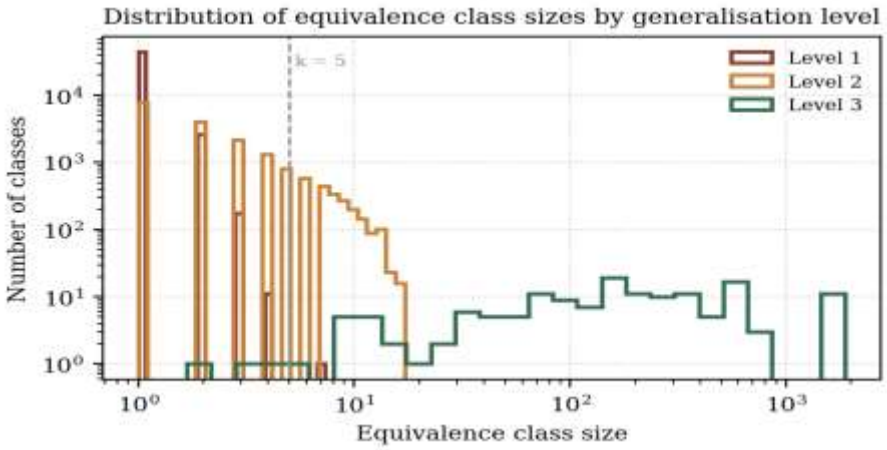

Figure 3. Distribution of equivalence class sizes at three generalisation levels, on logarithmic axes.

Figure 3 shows why a single k value is a poor summary. At L2 the smallest class is still one, so k-anonymity is 1 and the dataset fails any k-based rule; but the bulk of the distribution has moved right and only 15.7 % of records are unique. A release policy expressed purely as a k threshold treats L2 and L0 identically, when they differ by a factor of six in journalist risk.

3.3 Generalisation against suppression

Suppression — deleting records in classes below the target — is the alternative route to a k threshold. Table 6 and Figure 4 report its cost.

Table 6. Records suppressed (%) to reach a target k, by generalisation level

Generalisation level	k = 2	k = 5	k = 10	k = 20
L1	88.45	99.99	100.00	100.00
L2	15.71	55.26	86.70	100.00
L3	0.00	0.02	0.08	0.35

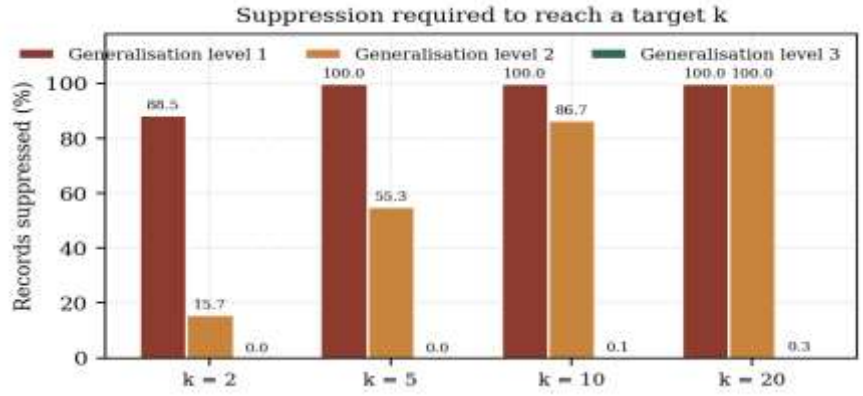

Figure 4. Proportion of records that must be suppressed to reach each target k, by generalisation level.

Suppression applied to a lightly generalised dataset is not a de-identification strategy but a destruction strategy. At L1, reaching k = 5 requires discarding 99.99 % of the cohort, leaving seven records. At L2, k = 10 costs 86.7 % of the data. At L3 the same threshold costs 0.08 %.

The practical rule that follows is to generalise first and suppress only the residue. Suppression is also not neutral in what it removes: the suppressed records are those in small classes, which are by construction the unusual patients — the rare diagnoses, the extreme ages, the small districts. A dataset made k-anonymous chiefly by suppression has been made anonymous by deleting the people least like everyone else, which biases every subsequent analysis in a direction that is hard to characterise and impossible to correct.

3.4 The risk–utility frontier

Applying the levels independently to the three variables gives 64 configurations. Figure 5 plots all of them.

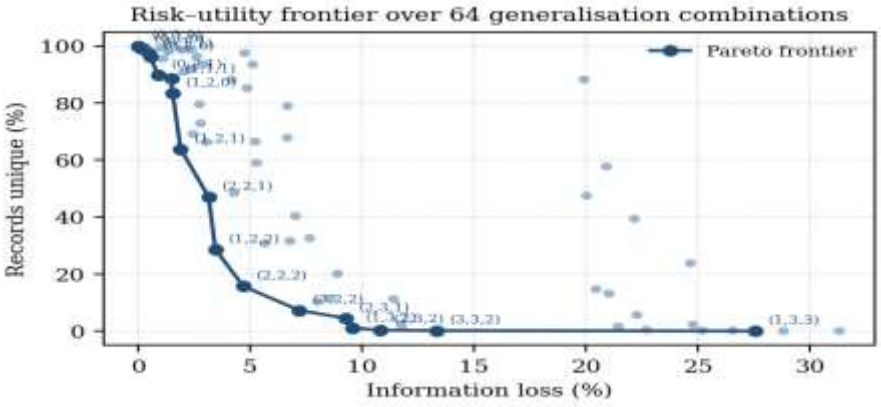

Figure 5. Risk against information loss for all 64 generalisation configurations, with the Pareto frontier marked. Labels give the (age, date, geography) level triple.

Most configurations are dominated: for the majority of points on the plot there exists another configuration that is both less risky and less lossy. The frontier is what a custodian should choose from, and it is short — a handful of configurations out of 64 — which is itself useful, because it reduces a large design space to a small menu.

The frontier also shows that the cheapest risk reductions come from coarsening the date and the geography rather than the age, which is the opposite of conventional practice: age banding is the intervention most often applied by default and, in this structure, the one that buys least.

3.5 Cohort Size Does Not Protect

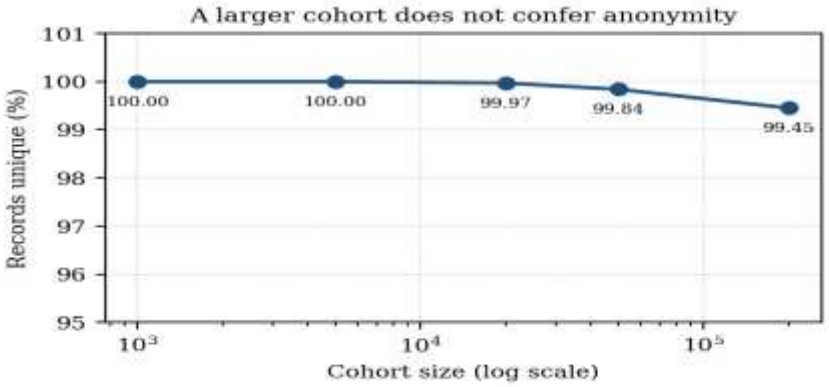

Figure 6. Proportion of unique records against cohort size, with no generalisation.

A common intuition holds that a record is safer in a larger dataset. Figure 6 shows that it is not: uniqueness on the full quasi-identifier set was 100 % at 1,000 records and 99.45 % at 200,000. Increasing n increases the number of distinct combinations available almost as fast as it increases the number of records, so the sparsity that creates uniqueness is preserved. Size is not a privacy control, and a release that would be unsafe at 1,000 records is not made safe at 200,000.

4. DISCUSSION

4.1 Findings and What a Custodian Should Do

Table 7. Findings and the corresponding practice

Finding	Practice
One high-cardinality variable took uniqueness from 14.6 % to 99.6 %	Compute risk per variable before review; prioritise by cardinality, not by how sensitive a variable feels
Exact dates were the dominant risk driver	Default to month or coarser unless the analysis demonstrably requires the day
Generalisation cut risk 300-fold with no measurable utility loss here	Measure downstream utility on the intended analysis rather than assuming a trade-off
Suppression to k = 5 cost up to 99.99 % of records	Generalise first; use suppression only for the residue
Suppression preferentially removes unusual patients	Report what was suppressed, by class, alongside the release
k = 1 at L2 despite 84 % of records being non-unique	Report the class-size distribution, not k alone
Uniqueness was ~100 % at every cohort size	Do not treat dataset size as a protection
Most of 64 configurations were dominated	Choose from the computed frontier, not from institutional habit

The second row deserves emphasis because it is the single highest-yield change available and is usually resisted by analysts as a matter of course. In this evaluation the exact admission date was worth 85 percentage points of uniqueness and contributed nothing detectable to the downstream model. Analysts should be asked to justify day-level dates rather than granted them by default, and the toolkit makes the cost of that default visible in a form that a data access committee can act on.

4.2 Relation to Existing Tools and Measures

Mature de-identification software exists and implements far more than this module does, including optimal generalisation search, differential privacy mechanisms and record linkage attack simulation. Nothing here competes with that. The contribution is a compact, readable implementation that a custodian can run and

audit in an afternoon, together with a worked demonstration of how the numbers behave, which is what is usually missing when a governance decision has to be made quickly.

Two limitations of the k-anonymity family apply to this toolkit as to any other. k-anonymity protects against identity disclosure and not against attribute disclosure: if every record in a class shares the same outcome, identifying the class identifies the outcome. The toolkit reports l-diversity for this reason, and in the evaluation l-diversity was 1 at every generalisation level, meaning that homogeneous classes existed throughout and attribute disclosure was never controlled. And none of these measures addresses an adversary with auxiliary information beyond the declared quasi-identifiers, which is the assumption under which high-dimensional data have repeatedly been shown to be re-identifiable in practice.

4.3 Limitations

The evaluation cohort is synthetic. As argued in Section 2.4, this is adequate for the risk results, which depend on combinatorial structure, and the structure used here — a skewed 220-category geography, a three-year daily date, five- and six-category categoricals — was chosen to resemble a routine extract. It is weaker support for the utility results, because the outcome model is synthetic too: the downstream AUROC of about 0.59 reflects a deliberately noisy generating process, and the finding that generalisation did not harm it may not transfer to an analysis whose signal lies in the generalised variables. The correct use of the toolkit is to run the utility evaluation on the intended real analysis, not to rely on the numbers reported here.

The generalisation hierarchies are fixed and hand-specified rather than searched. Optimal-search methods would find better points on the frontier of Figure 5, and the frontier reported is therefore a lower bound on what is achievable. Information loss is measured by cell width, which is one of several conventions and is not comparable across implementations.

The threat models assume the adversary knows exactly the declared quasi-identifiers and nothing else. Real adversaries have partial and noisy knowledge, which reduces risk, and auxiliary datasets, which increases it, and the net direction is not determined by anything computed here. Population uniqueness — whether a record unique in the sample is also unique in the population — is not estimated, and sample uniqueness overstates population risk for a sampled extract while understating it for a census.

Finally, the module addresses tabular data with declared quasi-identifiers. Free text, imaging, genomic data and longitudinal event sequences carry re-identification risks that this approach does not measure and should not be assumed to cover.

5. CONCLUSION

On a 50,000-record extract with a routine structure, removing direct identifiers left 99.85 % of records unique on five ordinary attributes, and almost all of that risk came from one variable: the exact admission date. Generalising the date to a year, the age to a 20-year band and the geography away eliminated unique records entirely and left the downstream model's discrimination unchanged. Reaching the same protection by suppression instead would have cost most of the dataset and would have removed the most unusual patients first.

The general lesson is that de-identification decisions currently made by convention can be made by computation, cheaply and before release. The specific lesson is that risk concentrates in high-cardinality variables that governance review tends not to focus on, and that the trade-off between privacy and utility is often much more favourable than custodians assume — but only because the variables that carry the risk and the variables that carry the signal are frequently not the same, which is a fact about a particular dataset and analysis and should be established rather than hoped for.

Declarations

Ethical approval: Not applicable.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of interes: The authors declare that there are no conflicts of interest regarding the publication of this paper.

Data availability: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author contributions:

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Bashar Sabah Sahib	✓					✓		✓			✓			✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

6. REFERENCES

1. L. Sweeney, "k-anonymity: a model for protecting privacy," Int J Uncertain Fuzziness Knowl Based Syst, vol. 10, no. 5, pp. 557-570, 2002.doi.org/10.1142/S0218488502001648
2. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: privacy beyond k-anonymity," ACM Trans Knowl Discov Data, vol. 1, no. 1, 2007.doi.org/10.1145/1217299.1217302
3. N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: privacy beyond k-anonymity and l-diversity," in Proc IEEE 23rd Int Conf Data Eng, 2007, pp. 106-115.doi.org/10.1109/ICDE.2007.367856
4. R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in Proc 21st Int Conf Data Eng. 2005, pp. 217-228.doi.org/10.1109/ICDE.2005.42
5. K. El Emam and F. K. Dankar, "Protecting privacy using k-anonymity," J Am Med Inform Assoc, vol. 15, no. 5, pp. 627-637, 2008.doi.org/10.1197/jamia.M2716
6. K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A systematic review of re-identification attacks on health data," PLoS One, vol. 6, no. 12, 2011.doi.org/10.1371/journal.pone.0028071
7. F. K. Dankar, K. El Emam, A. Neisa, and T. Roffey, "Estimating the re-identification risk of clinical data sets," BMC Med Inform Decis Mak, vol. 12, 2012.doi.org/10.1186/1472-6947-12-66
8. F. Prasser, J. Eicher, H. Spengler, R. Bild, and K. A. Kuhn, "Flexible data anonymization using ARX - current status and challenges ahead," Softw Pract Exp, vol. 50, no. 7, pp. 1277-1304, 2020. doi.org/10.1002/spe.2812
9. L. Rocher, J. M. Hendrickx, and Y. A. De Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," Nat Commun, vol. 10, 2019. doi.org/10.1038/s41467-019-10933-3
10. Y. A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: the privacy bounds of human mobility," Sci Rep, vol. 3, 2013. doi.org/10.1038/srep01376
11. A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in Proc IEEE Symp Secur Priv, 2008, pp. 111-125. doi.org/10.1109/SP.2008.33
12. C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Found Trends Theor Comput Sci, vol. 9, no. 3-4, pp. 211-407, 2014. doi.org/10.1561/0400000042
13. K. Benitez and B. Malin, "Evaluating re-identification risks with respect to the HIPAA privacy rule," J Am Med Inform Assoc, vol. 17, no. 2, pp. 169-177, 2010. doi.org/10.1136/jamia.2009.000026
14. P. A. Harris, R. Taylor, and R. Thielke, "Research electronic data capture (REDCap) - a metadata-driven methodology and workflow process for providing translational research informatics support," J Biomed Inform, vol. 42, no. 2, pp. 377-381, 2009. doi.org/10.1016/j.jbi.2008.08.010