




# Deident: An R package for data anonymization

Robert M. Cook <sup>1</sup>, Md Asaduzzaman <sup>2</sup>, and Sarahjane Jones <sup>1</sup>

<sup>1</sup> University of Staffordshire, Centre for Health Innovation, Blackheath Lane, Stafford, England 2  
University of Staffordshire, Stoke-on-Trent, England

DOI: [10.21105/joss.07157](https://doi.org/10.21105/joss.07157)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Susan Holmes](#)  

## Reviewers:

- [@PatrickRWright](#)
- [@nrennie](#)

Submitted: 06 August 2024

Published: 17 January 2025

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## Summary

The delivery of quality health care is a constant act of balancing demand against capacity, with emerging, data intensive, artificial intelligence (AI) and machine learning (ML) approaches poised to bridge gaps in the large, resource-limited sector ([Harwich & Laycock, 2018](#); [Nelson et al., 2019](#); [Wilson, 2019](#); [Yu et al., 2018](#)). The scale of data required in such projects magnifies the importance of existing ethical and legal frameworks for research with human participants ([Sales & Folkman, 2000](#); [UKRI, 2022](#)), notably around the risks posed by the processing of personally identifiable data (PID) and pseudo-PID (variables which if used together can identify an individual) ([ICO, 2023](#)).

One approach to dealing with PID concerns is to apply transformations to the data, e.g. encryption of names, or aggregation of ages, which can limit the risk of identification at the cost of nuance ([Tachepun & Thammaboosadee, 2020](#)). Hence, we demonstrate an extendable package of tools for the implementation and application of deidentification techniques to panel data sets.

## Statement of need

In order to broaden the access to sensitive data, data handlers need an auditable open process for the application of data transforms that limit the risk of personal identification (sometimes referred to as 'data masking' transformations). Several production scale systems exist for applying data masking transformations (e.g. [Delphix](#), [K2View](#), and [Accutiv](#)), but these are often expensive and cumbersome, limiting uptake in the domain of research. The research community hence requires a simple toolbox for the application of several common 'data masking transforms' implemented in an open source language.

This implementation of the deident methods is in R, chosen due to the increased adoption of open source software into the working practices of the NHS (driven by the growth on the NHS-R community and similar groups). However, the underlying specification, and implementation of the transforms can be ported to other languages.

This package was designed considering the ICO guidelines on anonymization ([ICO, 2012](#)) and draws on pre-existing methodologies and naming conventions ([Garfinkel, 2015](#); [Integrate.IO, n.d.](#)). The package implemented de-identification via:

- pseudonymization – the consistent replacement of a string by a random string
- encryption – the consistent replacement of a string by an alpha-numeric hash using an encryption key and salt
- shuffling – replacement of columns by a random sample without replacement
- blurring – the aggregation of numeric or categorical data according to specified rules
- perturbation – the addition of user-defined random noise to a numeric variable

Following the design principles of the existing tidyverse ([Wickham et al., 2019](#)) domain for

ease of adoption. The package includes tools to create a single pipeline for the application of a multi-step deidentification pipeline to multiple files stored within the same repository, alongside the option to serialize/ define the pipeline to/ via yaml. This approach allows a researcher to design and implement an appropriate de-identification plan and deliver it to the research support/ business intelligence team of an organisation with limited knowledge of the sensitive data. The supply of an easy method for de-identifying data sets which requires little scripting knowledge by Trust staff may aid in overcoming several information governance risks that keep operational data siloed within health Trusts.

## Comparison to existing R packages

Several packages have undergone development for the implementation of encryption methods to minimize identifiability within data sets (e.g. anonymizer (Hendricks, 2017), deidentifyr (Wilkins, 2019) and digest (Antoine Lucas et al., 2021)). While these packages implement a variety of encryption tools, such systems are not infallible (Szikora & Lazányi, 2022; Wang & Yu, 2005) especially if the data being encrypted is drawn from a known domain such as common names. This package introduces the stateful 'pseudonymization' method by which string vectors are replaced by a randomly generated hash the first time they are observed and then preserved for re-use. As such, breaking a single hash no longer exposes the 'key' and 'salt' the encryption relies on. In addition, this package introduces methods for removing pseudo-identifiability (identification via the combination of features) via the other methods included.

## In practice

To install the current version of the deident package, run the command

```
# install.packages("pak")
pak::pkg_install("Stat-Cook/deident")
```

The core functionality of the package is the deident function. To demonstrate functionality we use a subset of the babynames data set (Wickham, 2021) consisting of the final two years.

```
babynames <- babynames::babynames |>
  dplyr::filter(year > 2015)
```

```
str(babynames)
```

```
## tibble [65,448 × 5] (S3: tbl_df/tbl/data.frame)
## $ year: num [1:65448] 2016 2016 2016 2016 2016 ...
## $ sex : chr [1:65448] "F" "F" "F" "F" ...
## $ name: chr [1:65448] "Emma" "Olivia" "Ava" "Sophia" ...
## $ n : int [1:65448] 19471 19327 16283 16112 14772 14415 13080 11747 10957 10773 ..
## $ prop: num [1:65448] 0.0101 0.01002 0.00844 0.00835 0.00766 ...
```

The deident function produces a pipeline of actions and the variables to be transformed, with each subsequent call of deident adding a further transformation. The first argument of deident can either be a data frame or the output of a previous deident call.

The simplest use case is transforming one variable via a single method:

```
library(deident)

pipeline <- deident(babynames, "psudonymize", name)

apply_deident(babynames, pipeline)
```

The same method can be applied to multiple variables by adding the variable names:

```
pipeline2 <- deident(babynames, "psudonymize", name, sex)
```

```
apply_deident(babynames, pipeline2)
```

The transformations can also be initialized as base classes, allowing for them to be shared between pipelines:

```
psu <- Pseudonymizer$new()
```

```
pipeline3 <- deident(psu, Var1, Var2)
```

```
pipeline4 <- deident(psu, Var3, Var4)
```

Performing a multi-stage transformation can be done by chaining together calls to `deident`

```
blur <- NumericBlurrer$new(cuts=c(0, 10, 100, 1000, 10000))
```

```
pipeline3 <- deident(babynames, "psudonymize", name, sex) |>  
  deident(blur, n)
```

```
apply_deident(babynames, pipeline3)
```

An in depth example can be found in the `Worked Example vignette`, while more details on each method are presented in the `deident transforms vignette`.

## Current applications

The `deident` toolbox was developed for applications in the NuRS and AmreS research projects which aim to extract and analyze retrospective operational data from NHS Trusts to understand staff retention and patient safety.

## Contributions

The package was designed by RC, MA, and SJ. Implementation was done by RC. Quality assurance was done by MA. Documentation was written by RC, and SJ. Funding for the work was won by RC and SJ.

## Acknowledgements

The development of `deident` was part of the he NuRS and AmReS projects funded by the Health Foundation.

## References

- Antoine Lucas, D. E. with contributions by, Tuszynski, J., Bengtsson, H., Urbanek, S., Frasca, M., Lewis, B., Stokely, M., Muehleisen, H., Murdoch, D., Hester, J., Wu, W., Kou, Q., Onkelinx, T., Lang, M., Simko, V., Hornik, K., Neal, R., Bell, K., de Queljoe, M., ... Winston Chang., and. (2021). *Digest: Create compact hash digests of r objects*. <https://doi.org/10.32614/cran.package.digest>
- Garfinkel, S. L. (2015). *De-identification of personal information*. NIST. <https://doi.org/10.6028/NIST.IR.8053>
- Harwich, E., & Laycock, K. (2018). Thinking on its own: AI in the NHS. *Reform Research Trust*.
- Hendricks, P. (2017). *Anonymizer: Anonymize data containing personally identifiable information (PII) in r*. <https://github.com/paulhendricks/anonymizer>

- ICO. (2012). *Anonymisation: Managing data protection risk code of practice*.
- ICO. (2023). *What is personal information: A guide*. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-information-a-guide>
- Integrate.IO. (n.d.). *6 steps for data pseudonymization*. <https://www.integrate.io/blog/6-steps-to-pseudonymize-pii/#steps>
- Nelson, A., Herron, D., Rees, G., & Nachev, P. (2019). Predicting scheduled hospital attendance with artificial intelligence. *NPJ Digital Medicine*, 2(1), 26. <https://doi.org/10.1038/s41746-019-0103-3>
- Sales, B. D., & Folkman, S. E. (2000). *Ethics in research with human participants*. American Psychological Association.
- Szikora, P., & Lazányi, K. (2022). *The end of encryption?—the era of quantum computers* (pp. 61–72). Springer. [https://doi.org/10.1007/978-94-024-2174-3\\_5](https://doi.org/10.1007/978-94-024-2174-3_5)
- Tachepun, C., & Thammaboosadee, S. (2020). A data masking guideline for optimizing insights and privacy under GDPR compliance. *Proceedings of the 11th International Conference on Advances in Information Technology*, 1–9. <https://doi.org/10.1145/3406601.3406627>
- UKRI. (2022). *Framework for research ethics*. <https://www.ukri.org/councils/esrc/guidance-for-applicants/research-ethics-guidance/framework-for-research-ethics>
- Wang, X., & Yu, H. (2005). How to break MD5 and other hash functions. *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, 19–35. [https://doi.org/10.1007/11426639\\_2](https://doi.org/10.1007/11426639_2)
- Wickham, H. (2021). *Babynames: US baby names 1880-2017*. <https://doi.org/10.32614/cran.package.babynames>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemond, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Wilkins, D. (2019). *Deidentifyr*. <https://github.com/wilkox/deidentifyr>
- Wilson, C. (2019). *High-tech plans for the NHS*. Elsevier. [https://doi.org/10.1016/S0262-4079\(19\)31555-6](https://doi.org/10.1016/S0262-4079(19)31555-6)
- Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10), 719–731. <https://doi.org/10.1038/s41551-018-0305-z>