







# simulacrumWorkflowR: An R package for Streamlined Access and Analysis of the Simulacrum Cancer Dataset

Jakob Skelmose <sup>1,2</sup>, Lars Nielsen <sup>1</sup>, Jennifer Bartell <sup>2,3</sup>, Charles Vesteghem <sup>1</sup>, Martin Bøgsted <sup>1</sup>, and Rasmus Rask Kragh Jørgensen <sup>1,4</sup>

<sup>1</sup> Center for Clinical Data Science, Aalborg University, Aalborg, Denmark <sup>2</sup> Health Data Science Sandbox, University of Copenhagen, Copenhagen, Denmark <sup>3</sup> Center for Health Data Science, University of Copenhagen, Copenhagen, Denmark <sup>4</sup> Department of Hematology, Aalborg University Hospital, Aalborg, Denmark

DOI: [10.21105/joss.08120](https://doi.org/10.21105/joss.08120)

## Software

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

Editor: [Nick Golding](#) 

## Reviewers:

- [@tushardave26](#)
- [@aghaynes](#)
- [@drespresso](#)

Submitted: 27 February 2025

Published: 29 July 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 simulacrumWorkflowR package addresses the technical barriers associated with utilizing the Simulacrum through a streamlined workflow for accessing, preprocessing, and planning statistical analyses on the Simulacrum dataset. Our aim is to improve accessibility for researchers and clinicians with limited database expertise. The main function of this package is the `create_workflow()` function, which creates an R script based on the user's input that includes the necessary code for submission and execution on the United Kingdom's Cancer Analysis System (CAS) database servers ([Henson et al., 2020](#)).

## Statement of need

CAS data, stored in a restricted access Oracle database ([Oracle Corporation, 2025](#)), is held by the United Kingdom's National Disease Registration Service (NDRS) ([NDRS, 2023](#)). Researchers can analyze CAS data through assistance from NDRS staff and access to the Simulacrum, ([HDI, 2025b](#)) a synthetic version of the CAS database enabling the development and testing of SQL queries and analysis code. The latest version of the Simulacrum contains information about patient characteristics, tumor diagnosis, systematic anti-cancer treatment, radiotherapy, and genetic testing data ([Frayling & Jose, 2023](#)). Scripts developed using the Simulacrum can be sent to NDRS for execution on the CAS database; a common analysis workflow involves querying the database directly from R, extracting data, and further processing it to produce analytical outputs (the R workflow). NDRS staff adjust the code (e.g., SQL queries need further processing by NDRS due to 1) structural differences between the Simulacrum and CAS, 2) non-public details/specifications of the CAS database and 3) code alignment with NDRS best practices), execute the code, assess aggregated outputs with respect to patient privacy and return these outputs. If this process takes less than 3 hours, it's free of charge as of 2025. If the process is expected to take longer, the NDRS analyst may suggest simplifying the request or redirecting it to NHS England's data release services (DARS). For complex or recurring requests involving analysis, one can contact United Kingdom's Health Data Insight (HDI) ([HDI, 2025a](#)) for further assistance. HDI act as a specialized partner with the NDRS, who facilitate and executes the analysis on the CAS data as part of a collaborative framework ([Kafatos et al., 2025](#)). Thus, a streamlined process requires providing easily adaptable and executable code. The advantages of utilizing the Simulacrum can be summarized as follows:

1. Accelerated research.
2. Democratization of data.
3. Improving Privacy.

Accordingly, results from the Simulacrum should not be used to create real-world evidence - the dataset is intended for planning, designing, and testing analysis pipelines prior to generating actionable results with CAS data (Bullward et al., 2023). The process of having workflows executed on the real data through the Simulacrum requires users to construct SQL queries for extracting data from the CAS database and R code for analysis. To write and test SQL queries using the Simulacrum, users must download CSV files, install a local Oracle database and configure ODBC connections (Microsoft, 2024). SQL queries can then be executed from within an R script to extract data from the database for further analysis. Setting up a full Oracle database can be complex, particularly for new users. This presents a barrier to testing the full R workflow using the Simulacrum and may discourage users from doing so (NDRS, 2023). The simulacrum-WorkflowR package simplifies testing by removing the need to set up an Oracle database or configure ODBC connections. This allows users to create and test the full R workflow, including SQL queries that demonstrate the exact specification and form of the data required from the CAS database and how they integrate into the rest of the R script. This means NDRS can easily make the required adjustments before they are executed on the CAS database (Figure 1).

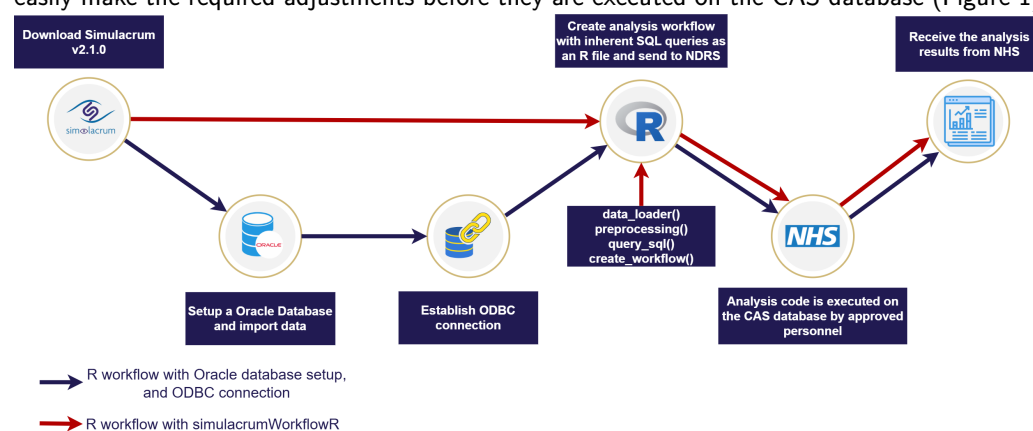


Figure 1: Flowchart of the process of running an analysis on the CAS Database using an R workflow tested on the Simulacrum and the process of running a similar analysis with the `simulacrumWorkflowR` package.

The `simulacrumWorkflowR` package is, to our knowledge, the first package designed to enhance usability and provide a complete workflow for utilizing the Simulacrum datasets to facilitate access and execution of code for analysis on the CAS database.

## Key functionalities

Providing a streamlined setup for building the workflow in R. The package includes:

- Integrated SQL Environment:** Leverages the `SQLdf` (Grothendieck, 2017) package to enable SQL queries directly within R, eliminating the need to set up an external database and ODBC connections by creating a local temporary SQLite database (SQLite, 2025) within the R environment.
- Query Helper:** Offers a collection of queries custom-made for the Simulacrum dataset for pulling and merging specific tables. Additionally, the `sqlite2oracle` function assists in translating queries to be compatible with the CAS database servers.
- Helper Tools:** Offers a range of data preprocessing functions for cleaning and preparing the data for analysis, ensuring data quality, and consistency. Key functions include cancer type grouping, survival outcomes, and logging reports.
- Workflow Generator:** Generates an R script with the complete workflow. This ensures correct layout and the ability to integrate the necessary code to obtain a workflow suitable for submission to NDRS and execution on the CAS database.

## Workflow illustration

simulacrumWorkflowR was developed with R version 4.3.3 (R Core Team, 2024). Installation requires Devtools and relies on dependencies listed in the DESCRIPTION file in the GitHub repository. These dependencies are automatically installed during package installation.

### Installation:

```
devtools::install_github("CLINDA-AAU/simulacrumWorkflowR",  
dependencies = TRUE, force = TRUE)
```

### Request to Download data:

```
open_simulacrum_request()
```

### Loading data:

```
library(simulacrumWorkflowR)  
  
dir <- "/path/to/simulacrum/csv/files"  
# Automated data loading  
data_frames_lists <- read_simulacrum(dir,  
selected_files = c("sim_av_patient", "sim_av_tumour"))
```

### Access dataframes individually

```
SIM_AV_PATIENT <- data_frames_lists$sim_av_patient  
SIM_AV_TUMOUR <- data_frames_lists$sim_av_tumour
```

### Querying data:

```
query_result <- "SELECT *  
FROM SIM_AV_PATIENT  
INNER JOIN SIM_AV_TUMOUR  
ON SIM_AV_PATIENT.patientid = SIM_AV_TUMOUR.patientid;"  
  
# Execute queries with the sql_test() function:  
df1 <- query_sql(query_result)
```

### Generating a reproducible workflow for submission to NHS

```
create_workflow(  
  libraries = "library(dplyr)  
              library(simulacrumWorkflowR)",  
  query = "select * from sim_av_tumour where age > 50 limit 500;",  
  data_management = "data <- cancer_grouping(query_result)",  
  analysis = "model = glm(AGE ~ STAGE_BEST + GRADE, data=data)",  
  model_results = "html_table_model(model)"  
)
```

### Oracle compatibility:

The `sqlite2oracle` function ensures basic query translation for Oracle databases.

## Limitations

**The Simulacrum Version:** The newest version of the Simulacrum is required for implementing the workflow on the CAS database because it resembles the CAS database more closely than earlier versions. Thus, the functionalities of this package are built for the Simulacrum v2.1.0, which means that some functions will not support earlier versions of the Simulacrum.

**Data Differences:**

- **Coverage:** The Simulacrum reflects diagnoses from 2016–2019, while CAS includes records dating back to 1971. The 2016–2019 restriction needs to be added to the code for running on CAS, as this time period is only provided in the free tier. Periods of the CAS data extending beyond The Simulacrum v2.1.0 will require a formal data release request and a cost estimate provided by DARS given the scope of data needed.
- **Structure:** The Simulacrum has a simplified structure for ease of use, but this differs from the evolving CAS database. Adjustment by NDRS is likely before execution on CAS.

The simulacrum, being a snapshot of a limited period and a simplified structure, inherently has structural and coverage limitations. This is because it is derived from a rather dynamic and complex original dataset. Despite these limitations, the Simulacrum offers a well-balanced comprehensive yet user-friendly test dataset.

**SQLite:** While both Oracle and SQLite use SQL syntax, there are notable differences between their syntaxes. For example, SQLite uses 'LIMIT' while Oracle uses 'ROWNUM'. The sqldf package's implementation also restricts table creation capabilities within SQLite. Adjustment by NDRS is likely before execution on CAS.

**Time Management:** While the Simulacrum facilitates SQL query testing, time. Similarly, code adjustments will take time that is unaccounted for in the estimates provided by SimulacrumWorkflowR. Despite this limitation, the package remains useful for benchmarking other components of the R script and identifying performance bottlenecks.

## Acknowledgements

Jakob Skelmosé and Jennifer Bartell acknowledge funding by the Novo Nordisk Fonden (NNF20OC0063268) via the Health Data Science Sandbox (<https://hds-sandbox.github.io>). Martin Bøgsted and Rasmus Rask Kragh Jørgensen acknowledge funding by the Novo Nordisk Fonden (NNF23OC0083510) via the SE3D project (Synthetic health data: ethical deployment and dissemination via deep learning approaches). We greatly appreciate the feedback we received from Lora Frayling at Health Data Insight.

## References

- Bullward, A., Aljebreen, A., Coles, A., McInerney, C., & Johnson, O. (2023). *Research paper: Process mining and synthetic health data: Reflections and lessons learnt*. 468. [https://doi.org/10.1007/978-3-031-27815-0\\_25](https://doi.org/10.1007/978-3-031-27815-0_25)
- Frayling, L., & Jose, S. (2023). *Simulacrum v2 user guide*. Health Data Insight.
- Grothendieck, G. (2017). *Sqldf: Manipulate r data frames using SQL*. <https://CRAN.R-project.org/package=sqldf>
- HDI. (2025a). *Health data insight CIC (HDI)*. <https://healthdatainsight.org.uk/>.
- HDI. (2025b). *Simulacrum: Artificial patient-like cancer data to help researchers gain insights*. <https://simulacrum.healthdatainsight.org.uk/>.

- Henson, K. E., Elliss-Brookes, L., Coupland, V. H., Payne, E., Vernon, S., Rous, B., & Rashbass, J. (2020). Data resource profile: National cancer registration dataset in england. *International Journal of Epidemiology*, 49, 16–16h. <https://doi.org/10.1093/ije/dyz076>
- Kafatos, G., Levy, J., Jose, S., Frayling, L., McInerney, C. J., Moore, G., Johnson, O., Rashbass, J., & Rous, B. (2025). Leveraging synthetic data to facilitate research: A collaborative model for analyzing sensitive national cancer registry data in england. *Therapeutic Innovation & Regulatory Science*. <https://doi.org/10.1007/s43441-025-00820-z>
- Microsoft. (2024). *Microsoft open database connectivity (ODBC)*. <https://learn.microsoft.com/en-us/sql/odbc/microsoft-open-database-connectivity-odbc?view=sql-server-ver17>
- NDRS. (2023). *Guide to using the simulacrum to support NDRS data requests*. National Disease Registration Service (NDRS).
- Oracle Corporation. (2025). *Oracle*. <https://www.oracle.com/>
- R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://doi.org/10.32614/r.manuals>
- SQLite. (2025). *SQLite*. <https://www.sqlite.org/>