The Journal of Open Source Software

simulist: An R package to simulate disease outbreak line list and contacts data

Joshua W. Lambert [1,2¶], **Adam J. Kucharski** [1,2], **and Carmen Tamayo Cuartero** [1,2]

1 Department of Infectious Disease Epidemiology and Dynamics, London School of Hygiene & Tropical Medicine, London, United Kingdom **2** Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Fabian Scheipl

Reviewers:

- @mponce0
- @epiben

Submitted: 08 April 2025
Published: unpublished

Summary

Epidemic and pandemic preparedness and analytics response requires robust analysis methods and a deep understanding of outbreak data. We introduce {simulist}, an open-source R package for simulating realistic infectious disease outbreak data. It is designed to allow users to generate data with specific disease outbreak characteristics by enabling flexible parameterisation of processes and variables, in particular, infectious disease epidemiology (e.g. delay distributions), contact patterns and demographic information, to simulate two common datasets in outbreak settings: 1) line list data, and 2) contact tracing data. It also offers post-processing of line list data to replicate right-truncation of real-time outbreak data, as well as creating "messy" data with realistically incomplete and inconsistent values.

Statement of need

Synthetic data, which is generated by a known simulation process, is useful in many scenarios where a "ground truth" is required, including teaching concepts on data collection, data cleaning and data visualisation; and testing analysis methods for precision, adequacy and robustness. In the analysis of infectious disease outbreaks, termed *outbreak analytics* (Carter et al., 2021; Jombart, 2021; Polonsky et al., 2019), one of the most common forms of data is a line list. This is a tabular dataset where each row corresponds to a single case in the outbreak, recording information on person, place and time. Information for each case can include, but is not limited to, personal information (name, age, sex), key event dates (date of symptom onset, date of hospitalisation, date of outcome), clinical outcomes, and geographic information (coordinates, administrative area). Understanding and characterising the transmissibility and severity of infectious disease outbreaks typically use line list data (Cori et al., 2017).

In the ecosystem of R packages for epidemiology, there are a range of tools for simulating and analysing line list data (Jombart et al., 2025), however, there are no stable, actively maintained and tested R packages that simulate epidemiologically realistic line list data[1]. The {simulist}

[1] This was found from an unstructured scoping review we conducted of other outbreak simulation tools, documented here: https://epiverse-trace.github.io/simulist/dev/index.html#related-projects

R package offers functionality to simulate outbreak data (line list and contact tracing) with options for users to define the epidemiological and demographic characteristics of the outbreak, while also being designed on the principle of maximum interoperability with other related R packages, e.g. {epicontacts} (Campbell et al., 2017), {epiparameter} (Lambert et al., 2025), and {cleanepi} (Mané et al., 2024). Accepting epidemiological parameters directly from {epiparameter}, an R package with a library of parameter estimates from the literature, facilitates generating synthetic data with the characteristics of past outbreaks.

State of the field

The majority of outbreak analytics uses R (R Core Team, 2025) and R packages extending the language to answer common epidemiological questions in an outbreak (Tamayo Cuartero et al., 2025). From basic data wrangling and data science using tools such as the tidyverse (Wickham et al., 2019), to using advanced methods, for example nowcasting/forecasting of transmissibility (Abbott et al., 2020) This is accompanied by several initiatives that teach epidemiology using R tooling (Batra et al., 2021; Valle Campos & Minter, 2025) resulting in an active ecosystem of open-source tooling, teaching, and community engagement (Morgan & Pebody, 2022), for example the epinowcast community.

However, to our knowledge, there are no existing tools that offer an easy method to generate synthetic outbreak data, with a defined and highly customisable simulation model. The {simulist} R package, part of the Epiverse-TRACE ecosystem (data.org, 2022), aims to fill this niche by offering line list simulation functionality that includes: custom parameterisation of epidemiological delay distribution, population-wide or age-stratified hospitalisation and death risks, uniform or age-structured populations, constant or time-varying case fatality risks, and customisable proportions of case types and contact tracing follow-up. This is complimented by post-processing to enable emulation of real-time outbreak dynamics (e.g. right-truncation and outbreak snapshotting) and messying data to mimic the types of issues encounter by people working with empirical outbreak data.

Key functions

{simulist} offers two modules of functionality: 1) simulation and 2) post-processing. In the simulation module there are three functions: `sim_linelist()`, `sim_contacts()` and `sim_outbreak()`. Respectively these simulate, an outbreak line list (`data.frame`), a contact tracing data set (`data.frame`), and lastly both an outbreak line list and contact tracing data set that are linked (i.e. from the same outbreak) (`list` of `data.frames`). All simulation function internally call a branching process model (Farrington, 2003) with the option of network adjusted contact patterns to account for the fact that choosing someone at random by following up a contact chooses individuals with a probability proportional to their number of contacts (Alexander & Day, 2010; Blumberg et al., 2014). All simulation functions can be run without requiring the user to specify any parameters (function arguments) (e.g. `linelist <- sim_linelist()`, see `?sim_linelist` for defaults), but all functions offer a range of arguments to modify the characteristics of the outbreak data. The parameters that can be specified by the user are organised into two levels. The parameters that control: the branching process, key epidemiological delays, risks, and demographic parameters are function arguments because they are more likely to be adjusted by users. Parameters thought to rarely require changing, except for advanced use cases are passed to the `config` argument, to prevent the function signature being overly complex. Parameters required by `config` can easily be generated and modified with the `create_config()` function.

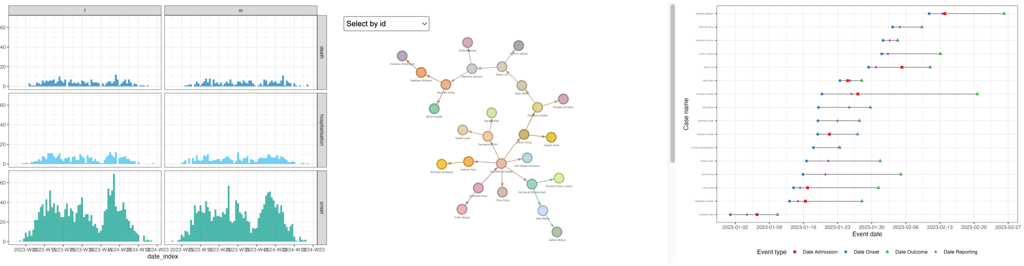

Figure 1: (a) daily incidence of symptom onset, hospital admission and deaths facetted by sex for an outbreak simulated with `sim_linelist()` and aggregated and plotted with `{incidence2}` (Taylor, 2020). (b) an `<epicontacts>` network plotted after converted from the output of `sim_outbreak()` (this network is interactive when rendered in an IDE or online). (c) a connected dot plot of the events in a line list simulated with `sim_linelist()`.

Here we highlight some of the simulation functionality available:

Age-stratified hospitalisation and death risks: the outbreak simulation includes three risks: hospitalisation risk (`hosp_risk` argument), death risk in hospital (`hosp_death_risk`) and death risk outside of hospital (`non_hosp_death_risk`), which all, by default accept a single number to define the population risk. In many infectious disease outbreaks risk is non-uniform across ages, so `{simulist}` accepts risks stratified by age using a user-defined `data.frame`, so any age-stratification is accepted. See the Age-stratified hospitalisation and death risks vignette for a complete explanation.

Age-structured population: the outbreak simulation in `{simulist}` assumes a uniform population age between a lower and upper age range (0 and 90 by default). However, in order to accurately generate synthetic data for outbreak scenarios where the population age structure is known this can be input into the outbreak simulation, again using a user-defined `data.frame` to flexibility specify an arbitrary number of age groups and proportions. See the Age-structured population vignette for more details and a worked example.

Time-varying case fatality risk: In addition to the ability to age-stratify hospitalisation and death risks, as outlined above, it is also possible to vary the fatality risk through time in the outbreak simulation. This is setup to accept a user-defined R function that takes the baseline risk and the simulation time (t) and calculates the fatality risk at time t. This is again setup in such a way as to maximise flexibility to the users and allows continuously varying (monotonic and non-monotonic functions)or discrete step-wise functions. The aim of this functionality is to phenomenologically mimic vaccination or non-pharmaceutical interventions without mechanistically modelling these in the simulation model. See the Time-varying case fatality risk vignette for more explanation and examples.

Simulating empirical outbreak data: The outbreak simulation produces exact, standardised, and complete data, hereafter referred to as *ideal* data. However, in reality those working with line list data often encounter a myriad of issues where the data they have available differs from the *ideal* data. In `{simulist}` we offer two post-processing functions that modify *ideal* line list data output by `sim_linelist()`to include two empirical characteristics, 1) real-time outbreak data and right-truncation (`truncate_linelist()`) and 2) inconsistent, missing, and duplicated data, so-called *messy* data (`messy_linelist()`).

Other functionality is documented in the Get Started vignette and Wrangling simulated outbreak data vignette.

Package Design

The `{simulist}` package is developed around certain design principles: *interoperability* with other R packages for upstream (e.g. `{epiparameter}`) or downstream (e.g. `{epicontacts}` ??) use; *ease of use* of functions (all simulation functions have sensible defaults for use out-of-the-box); and *transparency*, all design decisions are documented in the Design Principles vignette and help developer onboarding. Simulation functions output a `data.frame` (or `list` of `data.frame`s in the case of `sim_outbreak()`) enabling users to easily transform and manipulate the data, for example with the tidyverse R packages; and post-processing functions are compatible with `data.frame` subclasses (e.g. `<tibble>` or `<data.table>`).

Other resources

`{simulist}` is documented through vignettes, function documentation, it is also included within Epiverse-TRACE open-source tutorials.

Availability

The `{simulist}` R package is open-source and licensed under MIT. The source code is available on GitHub. The package can be downloaded from CRAN, R-universe or GitHub.

Acknowledgements

We would like to thank the entire Epiverse-TRACE team and community. Particular thanks to Hugo Gruson and Pratik Gupte for contributing and reviewing the package; Chris Hartgerink for reviewing the package; Sebastian Funk and James Azam for discussions about the simulation and real-time outbreak data; and Karim Mané, Abdoelnaser Degoot and Andree Valle Campos for discussions around data cleaning. Thanks to James Azam for providing useful feedback on the paper. Lastly, we thank the R-core team for the maintenance of the R language and the epidemiology R community. We thank data.org for their funding and support.

References

Abbott, S., Hellewell, J., Thompson, R. N., Sherratt, K., Gibbs, H. P., Bosse, N. I., Munday, J. D., Meakin, S., Doughty, E. L., Chun, J. Y., Chan, Y.-W. D., Finger, F., Campbell, P., Endo, A., Pearson, C. A. B., Gimma, A., Russell, T., CMMID COVID modelling group, Flasche, S., … Funk, S. (2020). Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Research*, *5*, 112. https://doi.org/10.12688/wellcomeopenres.16006.2

Alexander, H. K., & Day, T. (2010). Risk factors for the evolutionary emergence of pathogens. *Journal of The Royal Society Interface*, *7*(51), 1455–1474. https://doi.org/10.1098/rsif.2010.0123

Batra, N., Mousset, M., Spina, A., Florence, I., Liza, Aminatand, Laurenson-Schafer, H., Fischer, N., Molling, D., Polonsky, J., Bailey-C, Ebuajitti, Blomquist, P., Campbell, F., Hollis, S., Whoinfluenza, Llhaskins, Yurie, & Michellesloan. (2021). *The Epidemiologist R Handbook*. Zenodo. https://doi.org/10.5281/ZENODO.4752646

Blumberg, S., Funk, S., & Pulliam, J. R. C. (2014). Detecting Differential Transmissibilities That Affect the Size of Self-Limited Outbreaks. *PLoS Pathogens*, *10*(10), e1004452. https://doi.org/10.1371/journal.ppat.1004452

Campbell, F., Jombart, T., Randhawa, N., Sudre, B., Nagraj, V., Crellen, T., & Kamvar, Z. N. (2017). *Epicontacts: Handling, Visualisation and Analysis of Epidemiological Contacts* (p. 1.1.4). Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.epicontacts

Carter, S. E., Ahuka-Mundeke, S., Pfaffmann Zambruni, J., Navarro Colorado, C., Van Kleef, E., Lissouba, P., Meakin, S., Le Polain De Waroux, O., Jombart, T., Mossoko, M., Bulemfu Nkakirande, D., Esmail, M., Earle-Richardson, G., Degail, M.-A., Umutoni, C., Anoko, J.

N., & Gobat, N. (2021). How to improve outbreak response: A case study of integrated outbreak analytics from Ebola in Eastern Democratic Republic of the Congo. *BMJ Global Health*, *6*(8), e006736. https://doi.org/10.1136/bmjgh-2021-006736

Cori, A., Donnelly, C. A., Dorigatti, I., Ferguson, N. M., Fraser, C., Garske, T., Jombart, T., Nedjati-Gilani, G., Nouvellet, P., Riley, S., Van Kerkhove, M. D., Mills, H. L., & Blake, I. M. (2017). Key data for outbreak evaluation: Building on the Ebola experience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1721), 20160371. https://doi.org/10.1098/rstb.2016.0371

data.org. (2022). Epiverse. In *data.org*.

Farrington, C. P. (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*, *4*(2), 279–295. https://doi.org/10.1093/biostatistics/4.2.279

Jombart, T. (2021). Why development of outbreak analytics tools should be valued, supported, and funded. *The Lancet Infectious Diseases*, *21*(4), 458–459. https://doi.org/10.1016/S1473-3099(20)30996-8

Jombart, T., Rolland, M., & Gruson, H. (2025). *CRAN Task View: Epidemiology*.

Lambert, J. W., Kucharski, A., & Tamayo, C. (2025). *Epiparameter: Classes and Helper Functions for Working with Epidemiological Parameters* (p. 0.4.1). Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.epiparameter

Mané, K., Degoot, A., Ahadzie, B., Mohammed, N., & Bah, B. (2024). *Cleanepi: Clean and Standardize Epidemiological Data* (p. 1.1.0). Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.cleanepi

Morgan, O., & Pebody, R. (2022). The WHO Hub for Pandemic and Epidemic Intelligence; supporting better preparedness for future health emergencies. *Eurosurveillance*, *27*(20). https://doi.org/10.2807/1560-7917.ES.2022.27.20.2200385

Polonsky, J. A., Baidjoe, A., Kamvar, Z. N., Cori, A., Durski, K., Edmunds, W. J., Eggo, R. M., Funk, S., Kaiser, L., Keating, P., De Waroux, O. L. P., Marks, M., Moraga, P., Morgan, O., Nouvellet, P., Ratnayake, R., Roberts, C. H., Whitworth, J., & Jombart, T. (2019). Outbreak analytics: A developing data science for informing the response to emerging pathogens. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1776), 20180276. https://doi.org/10.1098/rstb.2018.0276

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Tamayo Cuartero, C., Carnegie, A. C., Cucunuba, Z. M., Cori, A., Hollis, S. M., Van Gaalen, R. D., Baidjoe, A. Y., Spina, A. F., Lees, J. A., Cauchemez, S., Santos, M., Umaña, J. D., Chen, C., Gruson, H., Gupte, P., Tsui, J., Shah, A. A., Millan, G. G., Quevedo, D. S., … Kucharski, A. J. (2025). From the 100 Day Mission to 100 lines of software development: How to improve early outbreak analytics. *The Lancet Digital Health*, *7*(2), e161–e166. https://doi.org/10.1016/S2589-7500(24)00218-8

Taylor, T. (2020). *Incidence2: Compute, Handle and Plot Incidence of Dated Events* (p. 2.6.1). Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.incidence2

Valle Campos, A., & Minter, A. (2025). *Epiverse-TRACE tutorials*.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686