Sharing
Tools and
Artefacts for
Reproducible
Simulations
in healthcare

Protocol for assessing the computational reproducibility of discrete-event simulation models on STARS

Amy Heather[1], Thomas Monks[1], Alison Harper[2], Navonil Mustafee[2], and Andrew Mayne[3]

[1]University of Exeter Medical School, Exeter, UK
[2]University of Exeter Business School, Exeter, UK
[3]Taunton and Somerset NHS Foundation Trust, UK

Publication date: 20th June 2024

This protocol will be used to assess the computational reproducibility of published healthcare discrete-event simulation (DES) models created using Python or R. It forms part of the project STARS: "Sharing Tools and Artefacts for Reproducible Simulations in healthcare".

This protocol is archived as a pre-registration. Haroz (2022)[1] identifies the Open Science Framework (OSF, `https://osf.io/`) and Zenodo (`https://zenodo.org/`) as suitable platforms for pre-registration.[1] In this case, Zenodo will be used as this is where other materials already exist for the STARS project, and so it can be stored alongside them in a Zenodo "community".

This work is supported by the Medical Research Council [grant number MR/Z503915/1].

Contents

1 Summary diagram

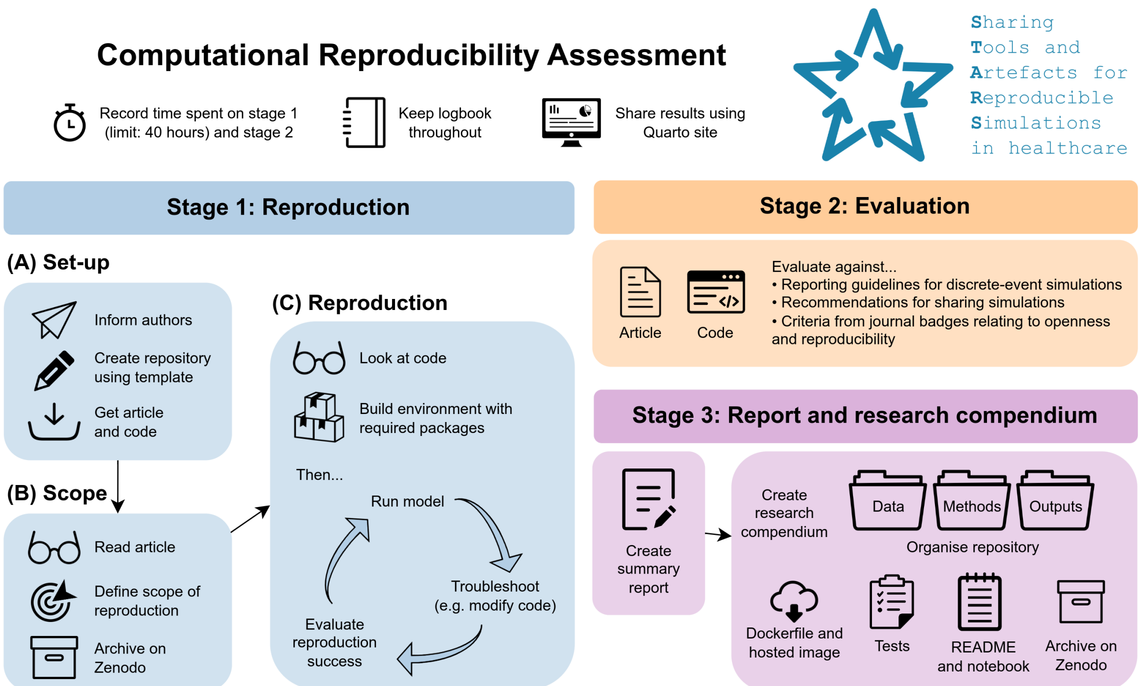

Figure 1: Workflow for assessing the computational reproducibility of discrete-event simulation models on STARS.

2 Prerequisites

This protocol ("Protocol for assessing the computational reproducibility of discrete-event simulation models on STARS") is shared under a CC BY 4.0 license (`https://creativecommons.org/licenses/by/4.0/deed.en`). It can be freely used and adapted (with attribution). Please be aware of the following prerequisites:

- Python[2] including typical methods for environment management like conda[3] and virtualenv[4]
- R[5] including typical methods for environment management like renv[6]
- Quarto[7]
- Docker[8]
- GitHub (`https://github.com/`)

3 Introduction

3.1 Computational reproducibility

This protocol is focused on **computational reproducibility**, which is defined as the ability to get consistent results with a prior study when using the same data and methods as that study.[9] Several prior studies assessing computational reproducibility have been used to inform the development of this protocol:

- Krafczyk et al. (2021):[10] Assessed the reproducibility of seven articles published in the Journal of Computational Physics.
- Wood et al. (2018):[11,12] Assessed the reproducibility of 109 published impact evaluations in low- and middle-income countries. Conducted in association with the replication programme of the International Initiative for Impact Evaluation (3ie).
- Schwander et al. (2021):[13] Assessed the reproducibility of four health economic obesity models.
- Laurinavichyute et al. (2022):[14] Assessed the reproducibility of 118 articles published in the Journal of Memory and Language.
- Konkol et al. (2019):[15] Assessed the reproducibility of 41 geoscientific articles from Copernicus Publications and the Journal of Statistical Software

It was also informed by:

- Ayllón et al. (2021):[16] Article with recommendations on keeping modelling notebooks to support completion of TRACE (TRAnsparent and Comprehensive model Evaluation) documents.
- The "Guide for Accelerating Computational Reproducibility (ACRe) in the Social Sciences":[17] Guidelines on conducting reproductions of published social science research.
- McManus et al. (2019):[18] Article proposing several possible definitions for success in reproducing or replicating models in health economics.
- Marwick et al. (2018):[19] Article recommending how to structure data analytical work as research compendiums using the R package structure.

3.2 Present study

This protocol has been developed as part of the project STARS: "Sharing Tools and Artefacts for Reproducible Simulations in healthcare". It will be used to assess the computational reproducibility of **six** published healthcare discrete-event simulation (DES) models. These will be selected from the models identified by Monks and Harper (2023).[20] The selection criteria are:

1. The model code is **publicly available**
2. The model is created using **Python and R**, as these are popular free and open-source software (FOSS) for the development of models like DES.[20] If possible, we will choose an equal split of three Python and three R models.

3. The code has an **open license** (either already published, or added upon request from the STARS team).

Throughout the study, results will be openly available and shared via a **Quarto website**.[7] This will compile information on the reproduction of the article. This includes the notebooks (.ipynb or .Rmd) producing the items in the scope, as well as a chronological log of work using Quarto blog posts, and then later, the summary report and detailed study results. The protocol will often refer to our **template repository** which can be viewed at `https://github.com/pythonhealthdatascience/stars_reproduction_template` and is archived on Zenodo (DOI:10.5281/zenodo.12168890).[21]

As in Figure 1, there are **three key stages** to this protocol which should be **worked through in the following order**:

1. Reproduction
2. Evaluation
3. Report and research compendium

Two important processes that will need to be **completed during some or all of the stages** are:

- Keeping a detailed record of work using a **logbook**
- **Timing** how long tasks take to complete

4 Logbook

A **logbook** should be kept throughout **all three stages** below. It provides a detailed record of work, recorded via **Quarto blog posts** using the template provided.[21]

As in the guidelines for keeping modelling notebooks from Ayllón et al. (2021),[16] the logbooks will consist of **daily**, dated, chronological entries. **Tags** will be used to help indicate the activity on each day, and enable posts to be filtered by activity.[16] Keeping a detailed log will support later understanding of what was done and support preparation of final documents like the summary report.

Each entry in the logbook should contain the following:

- **Researcher name** and **date**
- **Tags** (e.g. setup, scope, read, reproduce, guidelines, compendium, report)
- **Time** spent on tasks (if applicable)
- **Comprehensive** record of work. This should include a record of working through each stage in the protocol, detailing **successes** and any **issues** faced, any **solutions** found to problems, and any **changes made to the model code** (noting where and how the code was changed).
- Clear statement if and when each item in the scope is considered to have been **successfully reproduced**.

5 Timing

During the first and second stages of the study (**reproduction** and **evaluation**) we **time how long each task takes**. Whilst timing, the researcher should ensure to timestamp when they have:

- Finished reproducing each of the items in the scope
- Finished working on the evaluation of artefacts (badges and recommendations on sharing)
- Finished working on the evaluation of the article (reporting guidelines)

These times should be recorded within the logbook alongside each activity (e.g. 12:10 to 12:45). The times should be monitored with a **maximum of forty hours** allowed for the first stage (attempting to reproduce the study), as in Krafczyk et al. (2021).[10] This cut-off is implemented as we anticipate there would be little more to learn from spending longer than that time on reproducing a single study.

The only exceptions to this timing are:

- **Computation time**. This is at the researcher's discretion. For example, we suggest including short run times during which the researcher is continuously working on the study. However, we would suggest excluding longer run times during which the researcher is no longer working (such as setting the simulation to run for five minutes whilst they make a cup of tea).
- **Time spent by other researchers**. The recorded time should include time spent by the primary researcher on completing tasks and time spent in discussion with other researchers. However, if other researchers spend time preparing for those discussions (for example, reading the article or looking over the work), this does not need to be included.

6 First stage: reproduction

Remember! Record progress and the **time** spent on tasks in the **logbook** (as described in section 5).

6.1 Set-up

6.1.1 Inform the authors

This step may have already been completed (for example, if the author has already been contacted to request the addition of an open license to their repository). If not, the researcher should:

1. **Email the corresponding author** about the study, including a **copy of the study protocol**. A **template** for this email is provided in Appendix A.
2. If the email rebounds, search online for a **recent email address** for any of the study authors.

If the study authors do not reply, there is no need to follow up on this email asking for a response, as the email simply serves to notify the study authors and does not require a response.

6.1.2 Create repository using template

To set up the repository:

1. Go to `https://github.com/pythonhealthdatascience/stars_reproduction_template`[21] and select the "**Use this template**" button.
2. **Set up the repository** in 'pythonhealthdatascience' with the name 'stars-reproduce-surname-year' and description 'Assessing the computational reproducibility of surname et al. year as part of STARS.' Here, 'surname' and 'year' refer to the article being reproduced.
3. **Update the template** so that it is specific to the study being evaluated (e.g. titles, links, CITATION.cff).
4. Set up the provided **Python environment** ('requirements.txt') for creation of the Quarto site.

6.1.3 Upload code to the repository

We require that the model code has been shared under an open license. Hence, the researcher should be free to upload this to the repository. They should make a full copy of their code repository, uploading any **code** and related artefacts.

6.1.4 Update license for repository if required

Check which license the original study used for their code. By default, the template[21] includes an MIT license, but **this may need to be changed if the authors used a different license**, so it is compatible.

Following the guidance of the Turing Way[22] and R packages book,[23] the license does not need to be modified according to the packages used, unless code from that package is embedded within the work (for example, copy+and+paste a function) or if it is distributed as a binary with the work (i.e. bundled and stored with the work, rather than setting it to be exported from somewhere like PyPI or CRAN). If the code simply imports and uses functions from packages, this is not assumed to be derivative work, and hence a permissive license can be chosen. Likewise, the license can be permissive when using Docker[8] as we do not distribute the Docker image itself, but instead distribute the Dockerfile or refer to someone else to distribute (such as the GitHub container registry).[24]

6.1.5 Upload journal article to the repository

Before uploading the article to the repository, check which **license the article has been distributed under**. This is likely to be in a "copyright" section of the article. If it is distributed under a permissive license (e.g. CC-BY), then the researcher should be free to upload the article and artefacts to the repository. However, if the article does not have a license enabling reuse in this manner (which is likely if it is restricted access/behind a paywall), then they will likely be unable to. Although permission can be requested, this will often incur costs.

In these cases, the researcher should search for a **green open-access version** of the article. For example, copies of the article available on a pre-print site or in an institutional repository. If a green open-access article is identified, it can be uploaded to the repository - although it should be compared against the paywalled article to check for any differences between them before upload.

If an openly licensed version of the article cannot be located then **links to the original work** should be provided instead of uploading the article itself. If located though, the researcher should **upload all available materials** from the article to the 'original_study/' folder. This includes:

- The **journal article** and any **supplementary materials**
- Each **table and figure** from the main journal article as individual digital objects (e.g. .jpeg, .png) (not including those from the appendices/supplementary materials). These are uploaded to the repository separately as we will later use the files when describing the scope of the study on the Quarto site, and we will compare our results against them if they are within scope.

The researcher should amend the provided **template page** ('quarto_site/study_publication.qmd')[21] to display the PDFs or links to the journal article, and provide a link to the code. If using a green open-access article, a link to the paywalled article should also be included.

6.2 Scope of reproduction

6.2.1 Read the journal article

Read through the journal article (but not yet looking into the code or data). Any notes can be recorded within the logbook. These can be within a collapsible callout to aid readability of the log.

6.2.2 Define scope of reproduction

The next step is to define the scope of the reproducibility study - in other words, what parts of the paper we intend to reproduce. This should be focused on the **results of the simulation** (rather than other results like description of the sample). To identify the scope:

1. Look through each of the **tables and figures** in the article (excluding the supplementary material) and identify whether they are within scope or not (i.e. do they present results of the simulation).
2. Identify any **'key results' in the text** of the article that are within scope. These are results that are highlighted within the **abstract or results** section. This may include items from the supplementary materials if referred to in the text.
3. Evaluate whether the identified 'key results' are **already covered by the tables and figures** from the article. If not, they should be included in the scope.
4. Make a **consensus decision on scope with at least one other team member**.

Notes from thoughts and discussions around the potential scope can be recorded within the **logbook**.

The criteria for what is considered part of the scope is adapted from Wood et al. (2018).[12,11]

6.2.3 Compile items in scope

Once the scope has been decided, upload each item to the 'original_study/' folder. If the article does not have permissions to enable this, then the researcher can upload these materials to a separate **private repository** for their own reference, but cannot display these within the public repository.

- For **tables**, download a **CSV** version of each if available (or convert into CSV, if not).
- For **figures**, download the **highest-quality** version of each figure that is available (if not already saved in the repository).
- For results described in the **text** (but not captured in a table or figure), record in a format appropriate to then later compare against (for example, within a **CSV**).

The researcher should amend the provided **template page** ('evaluation/scope.qmd')[21] to display each of the items.

6.2.4 Archive scope on Zenodo

With the organisation linked to Zenodo, **toggle Zenodo to preserve** that repository, and then create a **release on GitHub**, which Zenodo will then automatically download and register with a DOI. The release should be recorded within the **changelog**.

This release serves as a public registration of the intended scope of the reproduction, archiving the repository at this point (prior to having started using or really looking at the code). As stated in the ACRe guide,[17] it is important that the scope is defined at the start of the study, and publicly archived so as not to be amended during the course of the study.[17]

6.3 Familiarise with artefacts

6.3.1 Look over the code/data

Browse through any code and data uploaded to the repository. The aim of this step is for the researcher to familiarise with the materials before setting up the environment or running the code. There are no requirements for how they should do this, but some suggestions include:

- In the logbook, record a one-sentence description of each file and tree of the uploaded files (as suggested by Ayllón et al. (2021)[16])
- Look for sections of code that produce items within the scope and record this within the logbook (as in Krafczyk et al. 2021[10]).

6.4 Set up environment

Identify the **software packages and their versions**, as used by the original study authors. This may be provided in an **environment file or within the article**. If not, the researcher should:

- Identify packages from those named to **import** within each of the scripts
- Select versions by looking at the **version history** for that package on a package repository (e.g. PyPI, CRAN), and identifying a version whose release date is **closest to but still prior to** the date of the code archive or paper publication (whichever is earliest).

Use or set up an environment file with the identified dependencies, and **create the environment**. Researchers should use simple methods for environment management such as Conda or VirtualEnv in Python, and renv in R. For simplicity, we are not requiring that they match the operating system used.

Important! This should be set up within the '**reproduction/**' folder, whilst the 'original_study/' folder should remain untouched.

6.5 Attempt to reproduce items in scope

This stage is an **iterative** process of running the code and attempting to reproduce the items in the scope. For this stage, the researcher should:

- Leave the 'original_study/' untouched - simply **copy over any relevant files** into the 'reproduction/' folder before running or modifying them.
- Use a **notebook** (.ipynb or .Rmd) when running the code (a "literate programming approach"), as this means the code and outputs can be shared easily. The notebook can be made available to view within the Quarto site by setting it as part of the toctree in '_quarto.yml'.
- Continue attempting to reproduce each item until they feel they have been **successful** (as defined in section 6.5.1) - or they run out of time (from the maximum forty hours allowed).
- **Troubleshoot issues**, contacting the study authors if necessary (as detailed in section 6.5.2)

Remember! Continue taking detailed notes and timings in the logbook, including each success and issue, and make note of any changes made to the model code. Whilst keeping notes in the logbook, it is recommended that **in-progress outputs are copied over** to the blog post folder (e.g. .png file for a figure), so that the researcher can easily and visually share progress within the log.

6.5.1 Successful reproduction

For **each item in the scope**, the researcher should decide whether it has been **successfully reproduced**. A binary decision should be made for each item (and none should be labelled as 'partial success').

This is a **subjective** decision. A successful reproduction does **not require that exactly the same results** are found. An item can be considered successfully reproduced if **minimal variation** is observed from the original results.

As an example, if it is possible to produce a table with some numbers being a match or very similar, but some numbers being substantially different, then this would be classed as having **not** been successfully reproduced. If however all aspects of the item were reproduced with reasonable similarity, this can be classed as **successful** reproduction.

Further recommendations:

- **Figures** - these are compared visually. Researchers should be unconcerned by **minor differences in presentation**.
- **Numbers** - researchers should calculate and report the **percentage difference** in results between the original study and those obtained by the researcher. As reported by Wood et al. (2018),[11,12] a meaningful difference in a value will vary between studies, and so it is difficult to set a single rule on what is or is not a minor difference. As such, researchers should follow a similar approach to Schwander et al. (2021),[13] considering whether the figure is reproducible at varying levels of percentage error (5%, 10% and 20%). However, they **should then use their judgement to decide whether the item has been reproduced**. This is similar to one of the definitions for "reproduction success" proposed by McManus et al. (2019)[18] - "*Results... vary only by XX% compared to the original, AND are consistent with the original conclusions*" - incorporating both numerical comparison and allowance for variability in whether this constitutes a meaningful difference from the original results.

Before concluding the reproduction, this decision should be run by at least one other researcher on the project, to ensure there is a **consensus decision** on whether the items were successfully reproduced. This conversation and the decision made should be included in the timing and recorded in the logbook.

In assessing reproduction success, it is important to note (as in Laurinavichyute et al. (2022)[14] and Wood et al. (2018)[11]) that the focus is **not** on the quality or robustness of the original results, or whether the main claims of the study are consistent. Instead, the focus is on whether it was possible to reproduce the article's results **within a reasonable margin of error** (given that we do expect a little variation, since discrete-event simulations are stochastic models, and may not have been fully controlled using random seeds or with any environment differences).

It is important to **clearly timestamp** in the logbook once the decision of "successful reproduction" has been made for each of the items in the scope.

6.5.2 Troubleshooting

Researchers should **troubleshoot** any issues encountered (including **making changes to the provided code**). In allowing modification and writing of code, our intention is that researchers try **as much as possible** to attempt to reproduce from the scope. The allowance of writing new code is similar to the approach of Krafczyk et al. (2021)[10] and the ACRe project.[17] Examples of changes they may need to make include:

- Correcting paths to files
- Correcting the versions of software, or adding missing packages or libraries
- Fixing errors in the code
- Adding code to produce an item in the scope, if not otherwise provided
- Adding a method for **controlling randomness** in the simulation (if not otherwise set up), so that they can get consistent results with themselves between re-runs of your notebook
- Adding a warm-up period (if suspected but not included in the code)

Troubleshooting can include asking **advice from other members of the STARS team**. In these cases, the researcher should ensure that they include a record the time spent, everything discussed, and any recommendations made.

6.5.3 Contacting the authors

Despite troubleshooting, the researcher may remain unable to run the code, or have large discrepancies with the original paper. In this case, once troubleshooting is exhausted, they should **contact the original author**. This email should:

- **Recap** the project (since our last email, when we informed them about the study).
- Link to the **Quarto site** with the documented reproduction attempt and list of issues that require resolution. Make sure the description of the problem is specific (e.g. identifying line in paper and place in code where we think something is missing, or where an issue is occurring).
- **Ask for suggestions** on an alternative course of action for issues, or for the complete code/data if missing.

If there is no response in two weeks, the researcher should contact them again. If there is still no response two weeks later, this can be marked as non-response. When emailing authors, it is suggested to follow the guidance on language and adapt from the **email templates** provided by ACRe in the chapter "Guidance for Constructive Communication Between Reproducers and Original Authors" from their guide.[17] The allowance of contacting authors is similar to the approaches of several studies,[10,11,17,25,15] with a maximum of four weeks for responses, as in Konkol et al. (2018).[15] This approach does however differ from Laurinavichyute et al. (2022)[14] who did not contact authors, since they considered reproducibility to be only about the available data and procedures and not anything shared privately.[14]

6.5.4 Running out of time

Once forty hours has passed, the reproduction stage stops and the researcher should move onto the evaluation stage for this paper. Before moving on, they should ensure:

1. If the model has no control for randomness (i.e. not implemented in original model and not yet added by researcher during troubleshooting stage) then they should select an alternative method for getting stable simulation results. This can be by doing a very large number of replications, or by assessing the required number of replications for a stable simulation.
2. Also, they should make a final decision on the reproduction success of each item.

6.6 Finishing up

6.6.1 Tidy up notebook and create reproduction success page

Tidy the reproduction notebook, so it simply produces each of the items in the scope, and clearly state how each section relates to the original article (e.g. captioning 'Reproduction attempt for Figure 2').

Using the **template page** ('evaluation/reproduction_success.qmd'),[21] show each item from the scope (as in the original article) alongside our best reproduction attempts (if possible under the article's license). Include the decision on the reproduction success for each item (along with any justification for this decision).

7 Second stage: evaluation

This section is completed **after** the attempted reproduction (so as to not interfere with timings).

Remember! Record progress in your **logbook** and **time spent** on each task.

Important: This evaluation is based on the **original** journal article or repository from the author (as in 'original_study/'), and not on the repository that was created whilst reproducing this study ('reproduction/'). If the original study had multiple repositories to choose from (e.g. development and archived code, both prior to publication date), remember to **refer to both of them** if there are any differences between them.

Getting a second opinion: If the researcher is uncertain about any criteria, they should note these in the **logbook**. Any criteria that were **unmet** or **uncertain** should then be discussed with at least one other researcher on the project to get a second opinion. Record the discussion (and its timing) in the logbook, and explain and justify the choices for uncertain items.

7.1 Badges

Several organisations and journals have developed badges which can be displayed alongside a research article to indicate how open and potentially reproducible it is, as detailed in Appendix B. These include the National Information Standards Organisation (NISO),[26] the Association for Computing Machinery (ACM),[26] the Institute of Electrical and Electronics Engineers (IEEE),[27] the Center for Open Science (COS),[28] and the journal Psychological Science.[29,30]

We will evaluate the original study artefacts (repository) against badges that relate to code (and not those specific to data), due to the nature of DES models (where "data" is often just parameters as part of the model script, with perhaps a few additional parameters in a separate data file within the repository). The **badges we will evaluate against** are:

- "Open objects" badges:
  - NISO "Open Research Objects (ORO)" and "Open Research Objects - All (ORO-A)"[31]
  - ACM "Artifacts Available"[26]
  - COS "Open Code"[28]
  - IEEE "Code Available"[27]
- "Object review" badges:
  - ACM "Artifacts Evaluated - Functional" and "Artifacts Evaluated - Reusable"[26]
  - IEEE "Code Reviewed"[27]
- "Reproduced" badges:
  - NISO "Results Reproduced (ROR-R)"[31]
  - ACM "Results Reproduced"[26]

  - IEEE "Code Reproducible"[27]
  - Psychological Science "Computational Reproducibility"[29,30]

The researcher should use the **provided template** (evaluation/badges.qmd)[21] to assess whether the artefacts from the original study meet the criteria for each of these badges. A **binary** decision is made for each criteria (as being either met or not met).

7.2 STARS framework

The artefacts (repository) associated with the original study will be evaluated against the **STARS framework**, which has essential and optional recommendations for sharing research artefacts from healthcare simulation studies. This framework was designed by Monks et al. (2024)[32] to complement and build on general open science recommendations from the Turing Way,[22] Taylor et al. (2017),[33] and the Open Modelling Foundation (OMF) minimal and ideal reusability standards.[34]

The researcher should use the **provided template** ('evaluation/artefacts.qmd')[21] to assess whether the artefacts from the original study meet the recommendations from this framework. Each criteria are evaluated as being "fully", "partially" or "not met".

7.3 Reporting guidelines

The **journal article** will be evaluated against two reporting guidelines for discrete-event simulation studies:

- STRESS-DES: Strengthening The Reporting of Empirical Simulation Studies (Discrete-Event Simulation) - Monks et al. (2019)[35]
- The generic reporting checklist for healthcare-related discrete event simulation studies derived from the International Society for Pharmacoeconomics and Outcomes Research Society for Medical Decision Making (ISPOR-SDM) Modeling Good Research Practices Task Force reports - Zhang et al. (2020)[36]

The researcher should use the **provided template** ('evaluation/reporting.qmd')[21] to assess whether the criteria from these guidelines are met by the journal article (including the supplementary material, although not including the code unless the article specifically refers to it for providing particular information). Each criteria are evaluated as being "fully", "partially" or "not met", with detailed evidence provided to support these claims (such as quotations from the article). If a criteria is not met by the original study, the researcher is welcome to make a suggestion in the evidence column of what they think the likely answer for that criteria might be.

8 Third stage: report and research compendium

8.1 Summary report

Use the **provided template** ('evaluation/reproduction_report.qmd')[21] to produce a simple summary report for the reproducibility assessment and evaluation. This template report includes:

- A short study description
- A citation to original study
- The number and percentage of items from the scope that were reproduced
- The time taken during reproduction and evaluated stages
- A description of the required troubleshooting steps
- Presentation of the reproduced items (e.g. figures, tables, statistics) alongside the items from the original study (or alongside description of and links to the original items, if they were not published under a license that allows reuse of the images).
- A percent stacked bar chart displaying the proportion of criteria met for each of the evaluations against guidelines

8.2 Research compendium

Once the computational reproduction has been completed, the repository will be restructured into a "**research compendium**". This is a term first introduced by Gentleman and Lang (2007)[37] which they defined as "*both a container for the different elements that make up the document and its computations (i.e. text, code, data,. . .), and as a means for distributing, managing and updating the collection.*"[37] In an article by Marwick et al. (2018),[19] which focuses on how to structure data analytical work, they recommend a research compendium as having three key components:

1. Files organised according to convention
2. Seperate data, methods and outputs
3. Specifying the environment used for the analysis.[19]

A research compendium might also be referred to as a "**reproducibility file bundle**",[38] or as a "**reproduction package**".[10] Although not required to be structured as a package, this can be helpful in providing a structure for dependency management and file organisation, and for continuous integration of automated code testing.[19]

For this study, the 'reproduction/' folder will be modified as per the proposed structure below. Some of these modifications may already be in place from having gone through the reproduction steps above. Our primary motivation in doing this step is to make it **easy and clear for someone to re-run** our reproduction, whilst making relatively **minimal changes** to the code itself. Hence, we are not necessarily amending the repository to meet all of the recommendations for best practice of sharing DES models.

8.2.1 Modify repository

Make the following changes to the **reproduction/** folder, if not already implemented:

- **Have separate folders for data, methods and outputs** - as recommended by Marwick et al. (2018).[19] The exception for this change is parameters coded into the scripts (since these would require a large amount of work and restructuring to separate from the scripts, contrary to our motivation in this stage).
- Create **tests** which check whether a user is able to get the same results as we obtained during the reproduction, based on comparison of CSV files.
- Create a **Dockerfile** and double-check it works (build image and run model notebook/s).
- Enable the GitHub action to publish the Docker image on the **GitHub container registry**.
- Make sure that **model notebook/s** contains:
  - Run time for the notebook.
  - A clear statement of which parts of the notebook produce each item from the scope.
- Ensure that the **README** contains:
  - A citation for the original study.
  - A simple summary of the model (potentially incorporating any diagrams of the model that were provided).
  - The scope of the reproduction (including images of the figures/tables from the original study).
  - An overview of the repository.
  - Instructions for setting up the environment.
  - Instructions for running the model (and reproducing items from the scope).
  - Instructions for running the test/s.
  - The hardware and software specs for the computer used for reproduction (machine, RAM, operating system and version).
  - The run time for the model.
  - Instructions for citation.
  - A short description of license.
- Ensure **Quarto site** displays the reproduction README and notebook/s.

8.2.2 Test-run with second team member

Once the research compendium is complete, a **second researcher** on the team should attempt to use it and confirm if they were able **reproduce** the results of the first researcher, and to check the compendium for **clarity**. This is similar to the approach of Krafczyk et al. (2021).[10] It should be recorded within the **logbook**.

8.3 Archive on Zenodo

Once modification of the repository is completed, a new GitHub release should be created to archive the repository on Zenodo (with record of this in the changelog).

8.4 Inform the authors

Email the authors again to:

- Thank the authors and let them know we have finished the assessment.
- Include a link to GitHub and Zenodo.
- Let them know how we are going to use the results from this work (i.e. lessons from reproduction, and guide framework design, and that this was not about validity of results).

A Appendix: Introductory email template

Subject: Reproducing results from [Title of paper]

Dear [Title (e.g. "Dr")] [Last name of corresponding author],

I am contacting you about your paper titled [Title] which was published in [Journal] in [Year] (vol [Volume], no. [Number]), [Link].

I am part of a research team working on a UKRI Medical Research Council funded project called Sharing Tools and Artefacts for Reproducible Simulations (STARS) in healthcare (`https://gtr.ukri.org/projects?ref=MR%2FZ503915%2F1`). We are writing to inform you that we were impressed with your paper and have selected it to be part of a STARS reproduction test.

STARS is a methods and training study investigating open science in healthcare simulation (specifically models written in Python or R). We aim to identify best practice and learn how we can support researchers make their models and code available, robust, reusable, reproducible, and preserved long term.

As part of STARS, we identified your paper as one of the small number of simulation studies where the code and model has been made available with an open license. The team would like to commend you for making your work open and for being a pioneer in this area.

The paper is one of several we have selected from the literature. Our initial study in STARS will attempt to run the published code and reproduce the results on our own computers. The work will record the challenges in installing, running and reproducing the results that are reported in the selected papers. As we are working with multiple studies our aim is to learn general lessons about sharing simulation models and move the field forward. This work is not about critiquing individual studies.

We have a full protocol describing the work available online: [Link to protocol on Zenodo]

We may contact you, if we are unable to run your model or obtain similar results. We appreciate that your time is valuable and we will only do so if we have exhausted all options. It is, of course, at your discretion if you reply to our query or not. We will also send you a copy of the results of the reproduction at the end of the study. The paper will be correctly cited and we will abide to the license conditions attached to the code.

There is nothing for you to do, and you do not need to reply to this email. However, do please feel free to contact us if you have any questions. Thank you again for making your work open so that others can learn from it.

[Reproducer full name]

[Reproducer job title], University of Exeter.

On behalf of the STARS team.

B Appendix: Badges

A badge is a label or image that is displayed alongside a published research article. There are currently several different badges from various organisations that can be used to indicate how open and potentially reproducible an article and its artefacts are. We have identified journal badges provided by the following organisations/journals:

- National Information Standards Organisation (NISO)[26]
- Association for Computing Machinery (ACM)[26]
- Institute of Electrical and Electronics Engineers (IEEE)[27]
- Center for Open Science (COS)[28]
- Springer Nature[39]
- Psychological Science[29,30]

Badges related to reproducibility are typically across three categories - "open objects", "object review" and "reproduced". However, there are other badges available that are related to reproducibility. These include badges for pre-registration of research, like the COS "Preregistered" badge.[28] There are also badges for replication, which is when an independent study on the same question find consistent results (potentially with new artefacts and methods). Examples of this are the NISO "Results Replicated (RER)" badge[31] and IEEE "Code Replicated" and "Dataset Replicated" badges.[27]

In some journals, these criteria are set as requirements for publication with the journal, rather than as badges. An example of this is the Psychological Science journal, which recently transitioned from awarding COS badges to making them requirements.[29]

We are focused on badges awarded by journals, but there are examples of badges that can be added to a repository if authors have followed a particular framework, either by self-allocating the badge or going through a review process. An example of this is Van Lissa et al. (2021)[40] who provide a package to facilitate use of a reproducible workflow in R, and suggest that a badge can be added to the README.md file of that repository if the package is used.[40]

B.1 "Open objects" badges

"Open objects" badges relate to permanently archiving digital objects in a public repository with a persistent identifier and open license.[31] Examples include:

- NISO "Open Research Objects (ORO)" and "ORO-A" (if all relevant objects available)[31]
- ACM "Artifacts Available"[26]
- COS "Open Data", "Open Materials" and "Open Code"[28]
- IEEE "Code Available" and "Datasets Available"[27]
- Springer Nature "Badge for Open Data"[39]

Table 1 compares the criteria for each of these badges (based on the sources cited above). As our focus is on DES models, we have excluded the badges that just relate to data, since simulation studies typically do not have "data", but instead have parameters (which often form part of the code itself). We have likewise excluded the COS "materials" badge since this is described as sharing the "components of the research methodology" - which in our case, would typically just be the article and the code.

The remaining badges either specifically relate to code, or are more broad: the NISO badge is relevant to "author-created digital objects used in the research (including data and code)",[31] and the ACM badge is relevant to "artifacts associated with the research".[26]

Table 1: "Open objects" badge criteria

Criteria	NISO	ACM	COS (code)	IEEE (code)
Stored in permanent archive that is publicly and openly accessible	✓	✓	✓	X
Has a persistent identifier	✓	✓	✓	X
Includes an open license	✓	X	✓	X
Complete set of materials are shared (as would be needed to fully reproduce article)	✓*	X	✓	✓
Artefacts are sufficiently documented, for researcher to understand how the code is used and relates to the reported methodology (e.g. package versions)	X	X	✓	X

* The standard "ORO" badge does not require this - but if all relevant research objects are available, the badge is modified to "ORO-A".

Abbreviations: ACM, Association for Computing Machinery; COS, Center for Open Science; IEEE, Institute of Electrical and Electronics Engineers; NISO, National Information Standards Organisation; ORO, Open Research Objects.

B.2 "Object review" badges

"Object review" badges relate to the digital objects (i.e. data, code) being reviewed according to the criteria of the badge issuer.[31] Examples include:

- NISO "Research Objects Reviewed (ROR)"[31]
- ACM "Artifacts Evaluated - Functional" and "Artifacts Evaluated - Reusable"[26]
- IEEE "Code Reviewed" and "Datasets Reviewed"[27]

Their criteria are summarised and compared in Table 2. The NISO badge is not included as it does not have criteria, but just states that badges in this category would evaluate against a specified set of criteria. The IEEE datasets badge is also excluded (as above for the "open objects" badges).

Table 2: "Open review" badge criteria

Criteria	ACM (functional)	ACM (reusable)	IEEE (code)
Artefacts are sufficiently documented, to enable them to be run	✓	✓	X
Artefacts are very carefully documented, to the extent that reuse is facilitated	X	✓	X
Artefacts are relevant to and contribute to the article's results	✓	✓	X
Complete set of materials are shared (as would be needed to fully reproduce article)	✓	✓	✓
Scripts can be successfully executed	✓	✓	✓
Artefacts are well structured, to the extent that reuse is facilitated, adhering to norms and standards of research community	X	✓	X

Abbreviations: ACM, Association for Computing Machinery; IEEE, Institute of Electrical and Electronics Engineers;

B.3 "Reproduced" badges

"Reproduced" badges are awarded when an independent party regenerates the article results using author objects.[31] Examples include:

- NISO "Results Reproduced (ROR-R)"[31]
- ACM "Results Reproduced"[26]
- IEEE "Code Reproducible" and "Dataset Reproducible"[27]
- Psychological Science "Computational Reproducibility"[29,30]

The criteria are summarised in Table 3. It should be noted that ACM specify that results are reproduced "in part" using artefacts from the author, and that exact reproduction is not required but that results should be within an acceptable range for experiments of that type.[26] Whether deviation in results or any modification of the author code (such as minor troubleshooting) is permissible is not detailed within the viewed criteria for the other badges.

Table 3: "Reproduced" badge criteria

Criteria	NISO	ACM	IEEE (code)	Psychological Science
Independent party regenerated results using the authors' research artefacts	✓	✓	✓	✓
Reproduced within approximately one hour (excluding compute time)	X	X	X	✓
Artefacts are well-organised	X	X	X	✓
Artefacts are clearly documented and accompanied by a README file with step-by-step instructions on how to reproduce results in the manuscript	X	X	X	✓

Abbreviations: ACM, Association for Computing Machinery; IEEE, Institute of Electrical and Electronics Engineers; NISO, National Information Standards Organisation.

References

[1] Steve Haroz. *Comparison of Preregistration Platforms*. Feb. 2022. DOI: https://doi.org/10.31222/osf.io/zry2u. URL: https://osf.io/preprints/metaarxiv/zry2u (visited on 05/10/2024).

[2] Python Core Team. *Python*. 2024. URL: https://www.python.org/.

[3] conda contributors. *conda: A system-level, binary package and environment manager running on all major operating systems and platforms.* July 2023. URL: https://docs.conda.io/projects/conda/.

[4] Python Packaging Authority. *virtualenv: Virtual Python Environment builder*. Dec. 2023. URL: https://virtualenv.pypa.io/en/latest/.

[5] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2024. URL: https://www.R-project.org/.

[6] Kevin Ushey and Hadley Wickham. *renv: Project Environments*. 2024. URL: https://rstudio.github.io/renv/.

[7] J. J. Allaire et al. *Quarto*. 10.5281/zenodo.5960048. Feb. 2024. URL: https://github.com/quarto-dev/quarto-cli.

[8] Dirk Merkel. "Docker: lightweight Linux containers for consistent development and deployment". In: *Linux Journal* 2014.239 (Mar. 2014), 2:2. ISSN: 1075-3583.

[9] National Academies of Sciences, Engineering, and Medicine et al. *Reproducibility and Replicability in Science*. eng. Washington (DC): National Academies Press (US), 2019. ISBN: 978-0-309-48616-3. URL: http://www.ncbi.nlm.nih.gov/books/NBK547537/ (visited on 06/18/2024).

[10] M. S. Krafczyk et al. "Learning from reproducing computational results: introducing three principles and the Reproduction Package". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2197 (Mar. 2021). Publisher: Royal Society, p. 20200069. DOI: 10.1098/rsta.2020.0069. URL: https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0069 (visited on 05/10/2024).

[11] Benjamin D. K. Wood, Rui Müller, and Annette N. Brown. "Push button replication: Is impact evaluation evidence for international development verifiable?" en. In: *PLOS ONE* 13.12 (Dec. 2018). Publisher: Public Library of Science, e0209416. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0209416. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209416 (visited on 05/10/2024).

[12] Benjamin Wood et al. "Replication Protocol for Push Button Replication (PBR)". en-us. In: *OSF* (Jan. 2018). Publisher: Open Science Framework. DOI: https://doi.org/10.17605/OSF.IO/YFBR8. URL: https://osf.io/yfbr8/ (visited on 05/10/2024).

[13] Björn Schwander et al. "Replication of Published Health Economic Obesity Models: Assessment of Facilitators, Hurdles and Reproduction Success". In: *Pharmacoeconomics* 39.4 (2021), pp. 433–446. ISSN: 1170-7690. DOI: 10.1007/s40273-021-01008-7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8009773/ (visited on 05/15/2024).

[14] Anna Laurinavichyute, Himanshu Yadav, and Shravan Vasishth. "Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy". In: *Journal of Memory and Language* 125 (Aug. 2022),

p. 104332. ISSN: 0749-596X. DOI: `10.1016/j.jml.2022.104332`. URL: `https://www.sciencedirect.com/science/article/pii/S0749596X22000195` (visited on 05/15/2024).

[15] Markus Konkol, Christian Kray, and Max Pfeiffer. "Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study". In: *International Journal of Geographical Information Science* 33.2 (Feb. 2019). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13658816.2018.1508687, pp. 408–429. ISSN: 1365-8816. DOI: `10.1080/13658816.2018.1508687`. URL: `https://doi.org/10.1080/13658816.2018.1508687` (visited on 05/15/2024).

[16] Daniel Ayllón et al. "Keeping modelling notebooks with TRACE: Good for you and good for environmental research and management support". In: *Environmental Modelling & Software* 136 (Feb. 2021), p. 104932. ISSN: 1364-8152. DOI: `10.1016/j.envsoft.2020.104932`. URL: `https://www.sciencedirect.com/science/article/pii/S1364815220309890` (visited on 05/13/2024).

[17] Berkeley Initiative for Transparency in the Social Sciences. *Guide for Advancing Computational Reproducibility in the Social Sciences*. Sept. 2022. URL: `https://bitss.github.io/ACRE/` (visited on 05/15/2024).

[18] Emma McManus, David Turner, and Tracey Sach. "Can You Repeat That? Exploring the Definition of a Successful Model Replication in Health Economics". en. In: *PharmacoEconomics* 37.11 (Nov. 2019), pp. 1371–1381. ISSN: 1179-2027. DOI: `10.1007/s40273-019-00836-y`. URL: `https://doi.org/10.1007/s40273-019-00836-y` (visited on 05/15/2024).

[19] Ben Marwick, Carl Boettiger, and Lincoln Mullen. "Packaging Data Analytical Work Reproducibly Using R (and Friends)". In: *The American Statistician* 72.1 (Jan. 2018). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2017.1375986, pp. 80–88. ISSN: 0003-1305. DOI: `10.1080/00031305.2017.1375986`. URL: `https://doi.org/10.1080/00031305.2017.1375986` (visited on 05/20/2024).

[20] Thomas Monks and Alison Harper. "Computer model and code sharing practices in healthcare discrete-event simulation: a systematic scoping review". In: *Journal of Simulation* 0.0 (2023). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17477778.2023.2260772, pp. 1–16. ISSN: 1747-7778. DOI: `10.1080/17477778.2023.2260772`. URL: `https://doi.org/10.1080/17477778.2023.2260772` (visited on 05/10/2024).

[21] Amy Heather et al. *Template for computational reproducibility assessments on STARS*. June 2024. URL: `https://zenodo.org/doi/10.5281/zenodo.12168890` (visited on 06/19/2024).

[22] The Turing Way Community. *The Turing Way: A handbook for reproducible, ethical and collaborative research (1.0.2)*. 2022. URL: `https://doi.org/10.5281/zenodo.7625728` (visited on 05/15/2024).

[23] Hadley Wickham and Jennifer Bryan. *12 Licensing*. Apr. 2023. URL: `https://r-pkgs.org/license.html` (visited on 05/21/2024).

[24] The Linux Foundation. *Docker containers: What are the open source licensing considerations?* URL: `https://www.linuxfoundation.org/resources/publications/docker-containers-what-are-the-open-source-licensing-considerations` (visited on 06/06/2024).

[25] Tom E. Hardwicke et al. "Analytic reproducibility in articles receiving open data badges at the journal Psychological Science: an observational study". In: *Royal Society Open Science* 8.1 (Jan. 2021). Publisher: Royal Society, p. 201494. DOI: `10.1098/rsos.201494`. URL: `https://royalsocietypublishing.org/doi/10.1098/rsos.201494` (visited on 05/15/2024).

[26] Association for Computing Machinery (ACM). *Artifact Review and Badging Version 1.1*. en. Aug. 2020. URL: https://www.acm.org/publications/policies/artifact-review-and-badging-current (visited on 05/20/2024).
[27] Institute of Electrical and Electronics Engineers (IEEE). *About Content in IEEE Xplore*. URL: https://ieeexplore.ieee.org/Xplorehelp/overview-of-ieee-xplore/about-content (visited on 05/20/2024).
[28] Ben B. Blohowiak et al. *Badges to Acknowledge Open Practices*. en. Publisher: OSF. Sept. 2023. URL: https://osf.io/tvyxz/ (visited on 05/20/2024).
[29] Tom E. Hardwicke and Simine Vazire. "Transparency Is Now the Default at Psychological Science". In: *Psychological Science* 0.0 (2023). DOI: https://doi.org/10.1177/09567976231221573.
[30] Association for Psychological Science (APS). *Psychological Science Submission Guidelines*. Dec. 2023. URL: https://www.psychologicalscience.org/publications/psychological_science/ps-submissions (visited on 05/20/2024).
[31] NISO Reproducibility Badging and Definitions Working Group. *Reproducibility Badging and Definitions*. en. Jan. 2021. DOI: 10.3789/niso-rp-31-2021. URL: https://www.niso.org/publications/rp-31-2021-badging.
[32] Thomas Monks, Alison Harper, and Navonil Mustafee. "Towards sharing tools and artefacts for reusable simulations in healthcare". In: *Journal of Simulation* 0.0 (2024). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17477778.2024.2347882, pp. 1–20. ISSN: 1747-7778. DOI: 10.1080/17477778.2024.2347882. URL: https://doi.org/10.1080/17477778.2024.2347882 (visited on 05/13/2024).
[33] Simon J. E. Taylor et al. "Open science: Approaches and benefits for modeling & simulation". In: *2017 Winter Simulation Conference (WSC)*. ISSN: 1558-4305. Dec. 2017, pp. 535–549. DOI: 10.1109/WSC.2017.8247813. URL: https://ieeexplore.ieee.org/document/8247813 (visited on 06/04/2024).
[34] The Open Modeling Foundation (OMF). *Reusability Standards*. OMF. May 2024. URL: https://www.openmodelingfoundation.org/standards/reusability/ (visited on 06/04/2024).
[35] Thomas Monks et al. "Strengthening the reporting of empirical simulation studies: Introducing the STRESS guidelines". In: *Journal of Simulation* 13.1 (Jan. 2019). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17477778.2018.1442155, pp. 55–67. ISSN: 1747-7778. DOI: 10.1080/17477778.2018.1442155. URL: https://doi.org/10.1080/17477778.2018.1442155 (visited on 05/21/2024).
[36] Xiange Zhang, Stefan K. Lhachimi, and Wolf H. Rogowski. "Reporting Quality of Discrete Event Simulations in Healthcare—Results From a Generic Reporting Checklist". In: *Value in Health* 23.4 (Apr. 2020), pp. 506–514. ISSN: 1098-3015. DOI: 10.1016/j.jval.2020.01.005. URL: https://www.sciencedirect.com/science/article/pii/S1098301520300401 (visited on 05/21/2024).
[37] Robert Gentleman and Duncan Temple Lang. "Statistical Analyses and Reproducible Research". In: *Journal of Computational and Graphical Statistics* 16.1 (Mar. 2007). Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/106186007X178663, pp. 1–23. ISSN: 1061-8600. DOI: 10.1198/106186007X178663. URL: https://doi.org/10.1198/106186007X178663 (visited on 05/17/2024).

[38] Florio Arguillas et al. "10 Things for Curating Reproducible and FAIR Research". eng. In: (June 2022). Publisher: Zenodo. DOI: https://doi.org/10.15497/RDA00074. URL: https://zenodo.org/records/6797657 (visited on 05/20/2024).

[39] Springer Nature. *Springer Nature Open data badge*. July 2018. URL: https://badgr.com/public/badges/xhW4FLHBRe6Tzz2Cj4Q1tA (visited on 05/20/2024).

[40] Caspar J. Van Lissa et al. "WORCS: A workflow for open reproducible code in science". en. In: *Data Science* 4.1 (Jan. 2021). Publisher: IOS Press, pp. 29–49. ISSN: 2451-8484. DOI: 10.3233/DS-210031. URL: https://content.iospress.com/articles/data-science/ds210031 (visited on 05/20/2024).