Amy Heather

University of Exeter Medical School

ORCID: https://orcid.org/0000-0002-6596-3479

Improving Reproducibility in Open Healthcare Simulation: Addressing Barriers and Promoting Best Practices

Submitted for the categories of Widening Reach & Improving Quality

In the

GW4 Open Research Prize 2025

Hosted by

The GW4 Alliance brings together the universities of: Bath, Bristol, Cardiff and Exeter. Find out more at www.gw4.ac.uk.

SUMMARY (100 word abstract)

This research generated actionable recommendations for sharing reproducible discrete event simulation (DES) models in healthcare. These were developed by systematically attempting to reproduce eight published healthcare DES models, the study identified barriers and facilitators to reproducibility. Open practices were used throughout: archiving, a pre-registered protocol, GitHub repositories and detailed logs for each reproduction, and results shared in an accessible format using Quarto websites hosted on GitHub pages. The recommendations generated – applicable to healthcare DES and broader computational research - can be used by researchers to make their work more reproducible, benefiting both the researchers themselves and the broader community.

What did you do?

My work aimed to provide **clear, actionable recommendations** on how to share **reproducible** DES models in healthcare. Reproducibility—the ability to achieve the same results using the same code and data—is vital for ensuring transparent, reliable and reusable research. DES is a method used to model dynamic systems where events occur at specific points in time, such as patients arriving at a clinic, waiting, and then receiving care.

To develop these recommendations, I systematically attempted to reproduce the results of eight published healthcare DES models in Python or R, all of which had shared code. For each model, I aimed to regenerate the tables, figures and key results as described in the corresponding research papers. Throughout this process, I documented the facilitators and barriers to successful reproduction. This empirical analysis was complemented by a thorough appraisal of whether these studies adhered to existing open research standards—such as DES-specific reporting guidelines and journal-specific open science badges (e.g., Association for Computing Machinery (ACM) artefacts for open code and data).

This work informed the development of a set of recommendations to support researchers in sharing reproducible models. While these guidelines are tailored to healthcare DES, many principles extend to computational research more broadly. The recommendations (Figure 1) provide practical steps to enhance reproducibility, such as providing code for all scenarios and sensitivity analyses; providing code to generate all the reported results, tables, and figures; and ensuring model parameters are correct in the shared code.

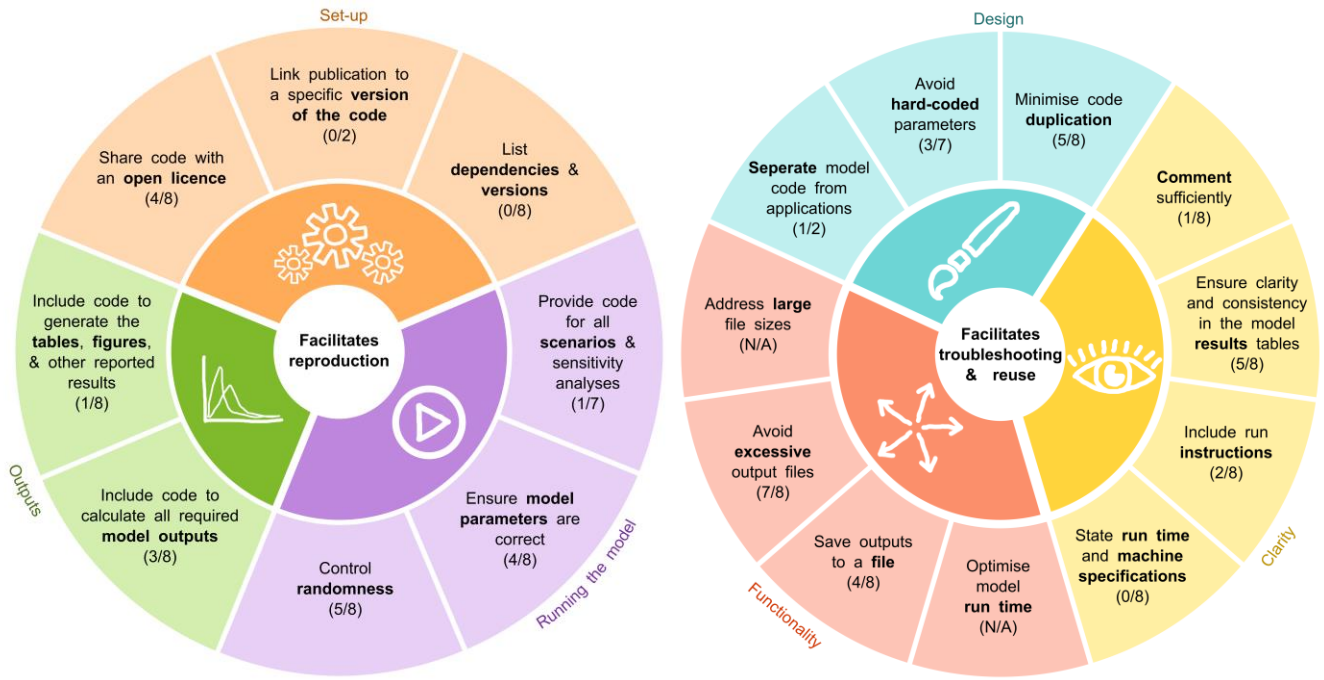

Figure 1. The recommendations generated from this research.

Why did you do it?

Nine out of ten publications describing healthcare DES models do not share the code for their model – but when they do, they often still lack the necessary artefacts and information to ensure reproducibility. Reproducibility is vital for re-running and verifying research findings, supporting model reuse, and maintaining trust in scientific outputs. It enables others to build on prior work and helps researcher revisit and build on their own analyses. Failures in reproducibility may indicate underlying issues, such as incorrect parameters, missing or outdated code, or unexpected changes in software behaviour due to updates. Troubleshooting non-reproducible code can be time-consuming or even impossible if required information is now lost.

A common barrier to ensuring shared models are reproducible is the perception that it requires advanced skills or significant time. This study aimed to address these challenges by generating a set of actionable recommendations to improve reproducibility. Several of these practices can be integrated throughout the research process with minimal additional effort. For example, using correct parameters, running scenarios, producing outputs, and creating tables and figures are standard parts of simulation research—sharing these complete artefacts is a straightforward but impactful step. Adding an open licence is another simple yet essential action that enables others to download, run, and use the code. While some recommendations require greater time investment, considering reproducibility from the outset helps to minimise these demands. The more practices researchers adopt, the greater the improvement in reproducibility.

How did you do it?

I wrote a **protocol** for the computational reproducibility assessments and evaluations, informed by existing studies which have attempted to reproduce computational results in other domains (e.g. physics). I conducted a

test-run of the paper on a model by a team member, attempting to reproduce their paper results, and then **pre-registered the protocol** on Zenodo (https://doi.org/10.5281/zenodo.12179845).

To ensure a consistent and transparent workflow, I created a **template GitHub repository** used for each reproduction attempt (https://doi.org/10.5281/zenodo.12168890). I systematically attempted to reproduce eight published healthcare DES studies in Python and R, using the authors' shared code and published materials. Throughout the process, I maintained **detailed logbooks** inspired by Ayllón et al. (2021)'s TRACE modelling notebooks, documenting time spent, actions taken, successes, failures, and every stage—from setting up environments to generating figures. Each reproduction was considered complete when either all outputs were successfully reproduced or further efforts were exhausted.

Each study underwent evaluation against two simulation reporting guidelines, the criteria for open journal badges (e.g., ACM badges), and a framework for sharing reusable simulation artefacts. I documented the facilitators and barriers encountered during these reproduction attempts, providing a comprehensive assessment of reproducibility challenges.

For every study, I created an **individual repository** containing the original code, modified code from the reproduction attempt, and a **Quarto website**. These sites included logbooks, evaluations, and reflections in an accessible format (example screenshot in Figure 2). All results were **public**, **archived** on Zenodo, and shared with the original authors for transparency and feedback.

I compiled the findings from all eight studies into a **summary website** (created using **Quarto and GitHub pages**), which consolidates the reproduction outcomes, evaluations, and reflections (https://pythonhealthdatascience.github.io/stars_wp1_summary/, screenshot in Figure 2). This work culminated in a clear set of actionable recommendations to improve the reproducibility of healthcare DES models, which I recently described in an **article** (currently in submission to a journal – but **pre-print** available at: https://doi.org/10.48550/arXiv.2501.13137).

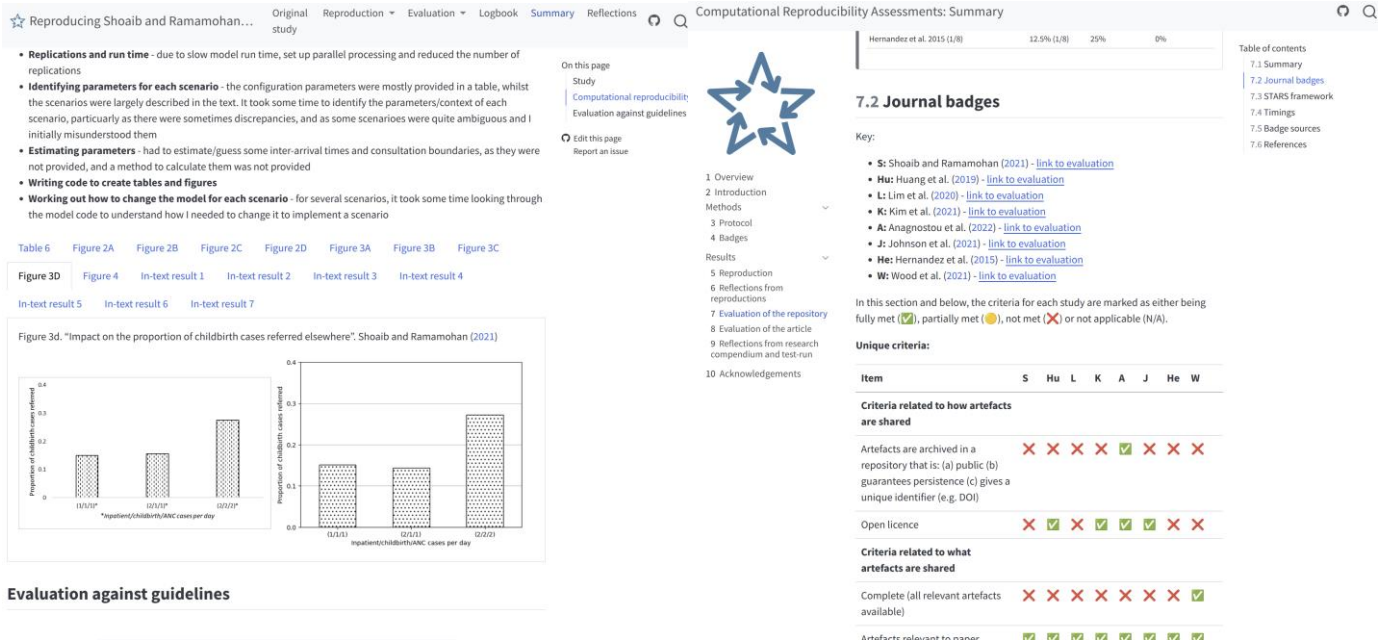

Figure 2. Example screenshots – left from a website describing the reproduction of a study, and right from the summary website.

What barriers / challenges did you have to overcome?

A major barrier was the extensive **troubleshooting** required due to missing or inconsistent parameters, incomplete scenario code, and the absence of open licenses.

Several models lacked documentation on software dependencies, making it difficult to recreate the computational environment and execute the code successfully. Authors of four of the eight articles assessed in this work had to be contacted to request the addition of open licences to their repositories.

Troubleshooting was time-intensive, with efforts taking up to 28 hours per paper (as in Figure 3). Despite these efforts, only half of the studies achieved full reproducibility. These challenges highlighted the practical difficulties researchers face when attempting to reproduce published models and underscored the need for complete code sharing, open licensing, and clearer documentation, to support reproducible research, among other things.

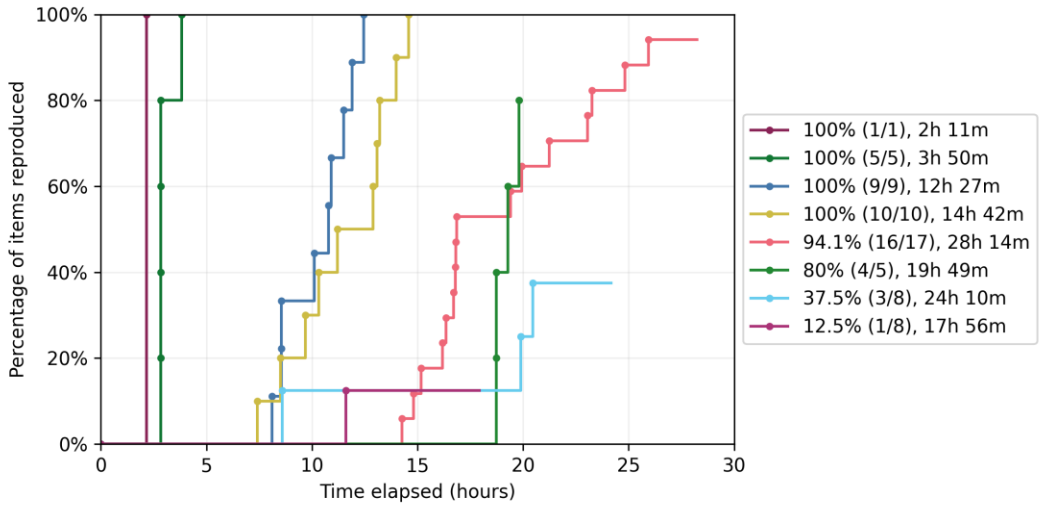

Figure 3. Count, proportion and time to reproduce items (tables, figures, reported results) in each of the studies.

What does it mean for you and your research/the research you support?

This work has three key applications for my current and future research:

1. **Developing reproducible template models:** I am currently creating template DES models in Python and R, hosted on GitHub, that follow a reproducible analytical pipeline (RAP). This approach, widely used in fields like public health and government analytics, ensures that analyses are scripted, version-controlled, and reproducible from raw data to final outputs. These templates will implement the recommendations from my findings - as well as the Levels of RAP framework developed by the NHS RAP Community of Practice. They will serve as practical examples to lower the barrier to entry for conducting reproducible DES research and provide a clear, replicable structure for others to follow.
2. **Refining the STARS framework:** The team I am part of previously developed the *Sharing Tools and Artefacts for Reusable Simulations* (STARS) framework, which provides guidelines for ensuring DES model reusability. Whilst STARS supports reuse, my work found it did not also ensure computational reproducibility. My findings will contribute to refining the framework to better address the practical challenges of reproducing healthcare DES models.

3. **Improving the STRESS-DES framework:** My supervisor developed the *Strengthening the reporting of empirical simulation studies-DES* (STRESS-DES) guidelines. These aim to facilitate replicability by enabling model reconstruction from a paper's description without requiring code. In my work, I found that adherence to the guidelines did not also translate to reproducibility. My findings will inform refinements to STRESS-DES that help incorporate practices necessary for computational reproducibility. Given its widespread adoption in healthcare modelling, these improvements will enhance the clarity and reproducibility of future studies following the STRESS-DES framework.

How might your findings / approach help other researchers?

My findings have practical implications across multiple areas of research:

1. **For modellers:** This work provides a structured framework for developing and sharing reproducible DES healthcare models – with broader applicability also to other computational research. By adopting practices such as open licensing, sharing complete code, and ensuring parameter accuracy, researchers can enhance the transparency, credibility, and long-term utility of their work.
2. **For peer reviewers:** Based on my findings, we offered suggestions in our article to help reviewers assess the likely reproducibility of submitted studies. This includes simple checks—like verifying open licenses, scenario code availability, and the provision of result-processing scripts—that can be integrated into the peer review process.
3. **For guidelines and frameworks:** This study highlights gaps in existing reporting frameworks like STRESS-DES and STARS. I have made suggestions to help improve clarity, encourage the sharing of complete materials, and better align with practical reproducibility needs.

We have plans to help share and disseminate these findings, including through the development of template models implementing the recommendations, applied examples, and collaborations with the Journal of Simulation and Health Data Research UK. At present, we have submitted an article describing the work (with pre-print available), and have disseminated the findings via the little book of DES (https://des.hsma.co.uk/), which is a resource publicly available and developed for attendees of the Health Service Modelling Associates Programme (HSMA), which is a programme of training and mentoring for people working in health, social care and policing in England.

Additional Information

I am an early career researcher (postdoctoral research associate) based at the University of Exeter. I was the primary researcher conducting this work, with supervision, support, involvement from, and thanks to: Tom Monks, Alison Harper, Nav Mustafee, and Andy Mayne.

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