OPEN-SOURCE MODELING FOR ORTHOPEDIC ELECTIVE CAPACITY PLANNING USING DISCRETE-EVENT SIMULATION

Alison Harper
Martin Pitt
Thomas Monks
PenCHORD, PenARC
University of Exeter Medical School
St Luke’s Campus, Heavitree Road
Exeter EX1 1TE, UNITED KINGDOM

ABSTRACT

The increase in elective surgical waiting lists as a result of the COVID-19 pandemic is creating significant consequences for health services worldwide. In the UK, the allocation of capital funds to increase capacity for managing elective waits has created planning and operational challenges for health services. This paper reports on the development and deployment of an interactive web-based discrete-event simulation model for supporting capacity planning of surgical activity and ward stay in a proposed new ring-fenced orthopedic facility in a UK health service. The model is free and open-source and developed to be generic and applicable for new capacity planning of elective recovery in orthopedics in other regions. With minor adaptations it can also be readily modified for application to other specialties. Given the current relevance of managing record elective waiting lists, there is potential widespread applicability of the simulation model which is supported by our open approach to modeling.

1 INTRODUCTION

Following the COVID-19 pandemic, growing elective waiting lists are one of the largest strategic problems facing healthcare services, with orthopedics representing one of the worst-performing specialties. The impact of the pandemic on orthopedic elective surgery wait times has been well documented both in England (Farrow et al. 2022) and worldwide (e.g., Momtaz et al. 2023). Increased time spent on waiting lists has been associated with poorer surgical outcomes, deterioration in physical function, and reduced quality of life (Kapstad et al. 2007; Vergara et al. 2011; Patten et al. 2022).

In the UK, National Health Service (NHS) access standards mandate providers to meet healthcare wait targets. Since 2012, it has been a statutory requirement that at least 92% of patients should have a referral-to-treatment time of less than 18 weeks. Elective wait times have been gradually increasing since this time, however from March 2020 after the onset of the COVID-19 pandemic, non-urgent elective activity was postponed to release capacity for inpatient and critical care. Since then, the waiting list for elective care has increased rapidly. In January 2023 there were over 7.2 million people on the waiting list, the highest number since records began. Of these, nearly 60% failed to meet the 92% standard of 18-week waits, and nearly 400,000 patients were waiting longer than 52 weeks (NHS England, 2023a).

In response to the long elective waits and reducing performance against core standards, NHS England (2023b) has allocated an extra £3.3 billion over 2023 - 2025, with a plan to eliminate waits longer than 65 weeks for elective care by March 2024 (NHS England, 2023b). This includes the allocation of capital funds to increase surgical capacity focusing on high-volume routine surgery. This ensures that emergency care
doesn’t disrupt operations by ring-fencing elective capacity to carry out low/medium complexity procedures at scale and are therefore fully resilient to winter pressures. These services are also in a position to improve efficiency of discharge processes as they are consistently treating and caring for the same specialties, and focus on standardizing care pathways (GIRFT, 2023). Trauma and orthopedics are one of six key priority specialties for the program.

This paper reports on the development of a free and open source discrete-event simulation (DES) tool for planning ring-fenced elective orthopedic capacity. The model is co-designed with a large NHS Trust in the Southwest of England, and is designed and licensed, to be reusable and adaptable, and to provide rapid information across a range of scenarios. It is also provided as a free interactive web app available to managers and clinicians to support elective recovery capacity planning.

2 LITERATURE REVIEW

Discrete-event simulation (DES) has been used for modeling elective waiting time dynamics, for example in radiology (Babashov et al. 2017); endoscopy (Carter et al. 2019); and urology (Chalk et al. 2019). For modeling post-COVID elective recovery, Wood (2022) simulated waiting list dynamics post COVID-19, finding that a temporary 25% increase in capacity over one year would restore performance to pre-COVID levels. Howlett and Wood (2022) likewise made the case for additional strategic investment in elective care resourcing at a national level. At a regional level, Nehme, Puchkova and Palikad (2022) used DES to predict the elective waiting list backlog and the delay in treatment based on a predetermined prioritization policy for endoscopy services, finding that a 55% capacity investment over the post-pandemic surge in waiting list levels was necessary to maintain the 92% NHS standard for performance. Specific to orthopedic services, modeling work undertaken by Mayne et al. (2022) found that without significant investment to increase capacity, combined with reconfiguration of services to establish efficient pathways, waiting times for joint replacement surgery will remain above 52 weeks across Ireland. These studies give a good indication of the magnitude of the problem facing healthcare providers after the pandemic but provide no information about how to plan services to meet this need, given the committed national investment to increasing capacity and tackle waiting lists at scale.

Operationally, several examples of orthopedic pathway models have been developed to improve patient flow, including detailed studies of individual orthopedic services (e.g., Rohleder et al. 2011; Baril et al. 2014). A generalizable, reusable knee surgery model was developed by Boyle and Mackay (2022), who found in their case study that adding three extra appointments per week can reduce the average waiting time by a third. A reusable model was developed by Suhaimi et al. (2018) for orthopedic clinics which allowed managers to modify parameters and identified internal efficiency improvements to reduce patient lengths of stay. Both models focused on flow through outpatient services.

For modeling theaters, a DES model addressed phased reopening strategies of theaters during COVID-19 across 18 surgical disciplines; the authors looked at bed and theater utilization according to length-of-stay scenarios, prior to full elective activity re-opening (Abdullah et al. 2022). Rachuba et al. (2022) showed how changes in weekly theater schedules affect theater utilization and overtime using DES which replicated daily processes in theaters and intensive care. Hassanzadeh et al. (2022) used DES to explore strategies for improving operating theater efficiency by exploring case-mix between emergency and elective care. Neither of these studies looked at ward bed requirements or the balance of capacity between theaters and wards for surgical throughput. A number of studies have demonstrated their generalizability and wider applicability, however even the models designed to be reusable and interactive are often not available for re-use by healthcare planners or researchers (e.g., Boyle & Mackay, 2022; Suhaimi et al. 2018; Hassanzadeh et al. 2022). A strength of this study is that the model is designed to be both immediately accessible and implementable, and available open source with licensing to enable it to be readily re-used or adapted to similar applications.
3 METHODS

We developed a DES model of surgical activity and ward stay in a proposed ring-fenced orthopedic service. The model was developed in collaboration with managers and clinicians in a health organization in England, serving a population of approximately 1 million people. In the early phases of the study, the model was planned to simulate orthopedic flow from both elective and non-elective sources to improve service efficiency in the face of rising emergency admissions and cancellations of planned care, particularly in the winter months. Subsequent to the pandemic, priorities shifted rapidly toward planning for efficient use of new capacity.

The study therefore had four aims:

- To develop an intuitive, interactive simulation tool that is flexible enough to answer a range of what-if questions;
- To make the tool available to clinicians and managers without the need to download and install software or to change and run code;
- To ensure that the model is simple enough to facilitate understanding and to be applicable more widely beyond the study site;
- To provide sufficiently accurate results and suitable output metrics to confidently support capacity planning of new elective hubs.

3.1 Model Objectives

We identified the components of the surgical inpatient process that were material to the questions being asked by our NHS collaborators (Figure 1). With the objective of maximizing surgical throughput, questions included:

- How many theaters do we need?
- How many beds do we need?
- What impact will reducing lengths-of-stay have on surgical throughput?
- What if fewer patients have a delayed length-of-stay waiting for out-of-hospital care?
- What if we run evening or weekend theater sessions?
- What if we change how we schedule surgery types, such that those with longer, more variable lengths-of-stay are scheduled earlier in the week?

These questions are all in line with high volume, low complexity surgical planning. Patients are likely to be less complex, with fewer co-morbid conditions and a high expectation that surgical and recovery processes will be uncomplicated. In turn, lengths-of-stay are likely to reduce, and discharge delays will be minimized with respect to ongoing out-of-hospital community care needs. In the partner NHS Trust, lengths-of-stay are national outliers, and a priority of the service is to reduce these in line with national benchmarks.

Output metrics of interest were:

- Bed utilization per day of week. Perhaps contrary to expectation, clinical staff indicated that bed utilization should be lower on weekends as there are fewer staff available. Mean bed utilization is outputted daily over model runtime, per weekday, and per experimental scenario with quantiles.
- Lost theater slots due to lack of beds. Theater slots may be lost for patient reasons such as illness, or for system reasons such as lack of bed availability. While medical patient ‘outliers’ will not
impact bed availability, the balance of beds to theater activity is a critical question as there is no buffer of an admission queue.

- Total surgical throughput. The primary goal of the high volume, low complexity hubs is to efficiently maximize surgical throughput, so the configuration which best achieves this - within other constraints relevant to service planners such as workforce availability - is a key output.

### 3.2 Model Data and Logic

Routinely collected data from an NHS Trust in England were used to identify patients receiving elective total joint replacement surgeries between January 2016 and December 2019. The Trust’s electronic health records (EHR) were used to identify elective total joint replacements using a combination of OPCS4 procedure and surgical site codes. Five core elective orthopedic surgical procedure types were identified and verified. A small number of short day-case ‘hip resurfacing’ surgeries [n=52] were removed from the dataset. The five remaining surgical types were classified into two classes: (I) **Primary** (primary hip replacement [p-THR n=3057; 51%], primary knee replacement [p-TKR; n=2302; 38%], uni-compartmental knee replacement [p-UKR; n=679; 11%]); (II) **Revision** (revision hip replacement [r-THR; n=482; 55%], revision knee replacement [r-TKR; n=392; 45%]). Each surgical procedure was fitted with a length-of-stay distribution to EHR data using Python package fitter (https://fitter.readthedocs.io/en/latest/).

Our model assumes an infinite waiting list. Baseline surgical theater scheduling rules define how patients enter the simulation model deterministically according to their surgical class. Theater scheduling rules were described and validated by two senior orthopedic surgeons and two theater scheduling managers. Baseline rules are:

- Three sessions per day, five days per week with no weekend activity.
- Morning and afternoon sessions schedule either: one Revision or two Primary surgeries
- Evening session schedules one Primary surgery

![Figure 1: Conceptual organization of orthopedic elective surgical activity and ward stay.](image)

Most patients did not remain in hospital once they were medically fit for discharge, however a proportion of patients in the EHR had a recorded *medically fit for discharge* date which preceded their actual discharge date (n=529; 7.6%). A length-of-stay distribution was fitted for all delayed discharge patient procedures. Figure 1 conceptualizes the organization of the service. Where patients are scheduled...
to arrive for surgery, but no bed is available within 0.5-1 day, the patient is flagged as a ‘lost slot’. This is an indication of a mismatch between demand (theater scheduling) and capacity (bed utilization). In reality, other system behaviors will account for some of these ‘lost slots’. For example, the slot may be lost for patient reasons such as illness; bed management activities may take place; or patients may be transferred back to the acute hospital for a variety of reasons, freeing up their bed.

Baseline parameters for the model were provided as: 40 beds and 4 theaters. Baseline length-of-stay parameters are in Table 1. In all cases, log normal distributions were used.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Mean (days)</th>
<th>Std (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-THR</td>
<td>4.4</td>
<td>2.9</td>
</tr>
<tr>
<td>p-TKR</td>
<td>4.7</td>
<td>2.8</td>
</tr>
<tr>
<td>p-UKR</td>
<td>2.9</td>
<td>2.1</td>
</tr>
<tr>
<td>r-THR</td>
<td>6.9</td>
<td>7.0</td>
</tr>
<tr>
<td>r-TKR</td>
<td>7.2</td>
<td>7.6</td>
</tr>
<tr>
<td>Delayed discharge</td>
<td>16.5</td>
<td>15.1</td>
</tr>
</tbody>
</table>

### 3.3 Open Modelling Framework

To facilitate use, reuse, modification, and refinement of the model, we adopted an open modeling approach (Harper & Monks, 2023). Interacting directly with the code is likely to be a barrier for managers and clinicians for whom the model is aimed. Additionally, while the UK NHS Python Community for Healthcare champion the use of Python and open code in the healthcare sector, our experience is that there are frequently obstacles to downloading and installing Python on NHS machines. In summary, our Open Modelling Framework supports modelers to share executable healthcare DES models in Python over the web. The framework is based on combining remote version control repositories with free and open-source software - Jupyter Notebooks, Jupyter Books, and Streamlit - along with free digital infrastructure provided by Binder, streamlit.io, and GitHub pages. It provides appropriate open licensing to enable reuse and adaptation of models and facilitates long-term archiving using open science repositories such as Zenodo. The framework enables executable models to be shared with users of different technical abilities: from coders to software literate users (Figure 2). The framework is flexible, and a user can adopt all or part of it to meet their needs.

A full description of the framework is provided in Harper and Monks (2023). Here we briefly describe four key elements: remote code repositories; licensing of research artifacts; Streamlit web applications; and long-term archiving in an open science repository.

**Remote code repository:** Central to Figure 2 is a remote code repository. The most popular solutions are GitHub, GitLab, and BitBucket. Most developers opt for GitHub, but we recommend modelers consider the unique advantages and disadvantages to each solution. All code artifacts (including license, meta-data files, and readme) should be committed to the remote repo to provide long lasting version control. Some developers store data here too, although other specialist solutions exist for data version control.

**Licensing of research artifacts:** Before sharing (or making any Python code public), it is essential that authors select an appropriate open license for the code and other research artifacts. This, for example, ensures appropriate use of the model and removes liability of the authors. A common approach in data science is to adopt a permissive license such as the MIT. Alternatively, depending on the software used, or preferences, a strong copyleft license could be adopted to ensure future releases/adaptations for the model are released under the same license. An example is the GNU General Public License v3 (GPL-3). See Monks, Harper, Anagnostou, and Taylor (2022) for more details on license type.
Streamlit web applications: A web app interface to the model is suitable for less technical simulation users such as NHS analysts with no coding experience, NHS managers and clinicians, researchers not familiar with Python, or the general public. A web app will provide simple ways to setup and execute the model. This might include parameter fields, logic diagrams, basic animation, and buttons. Streamlit is a simple modern Python library that provides a way to script web applications. Streamlit has many built in controls and display functions, for example, buttons, sliders, text fields, tables, and charts. Using Streamlit, it is possible to create a very simple simulation app with only a few lines of code; for example, basic parameter settings, an execute button, and results displayed as a table in the browser. Streamlit.io is a free hosting service for Streamlit apps. When used together they provide a robust and easy way to share simulation models to healthcare users. These models can then be accessed from any device. For example, models can be executed from a phone or a tablet instead of traditional laptop or desktop environment.

Open science repository: Lastly, for long term archiving (and preservation) of simulation models our framework links code in the remote repository to an open science repository. Examples, include Figshare, Zenodo, and the Open Science Framework. Each of these has guarantees on storage and mints a Digital Object Identifier (DOI) to enable citation and improve discoverability.

Figure 2: Open Modelling Framework (reproduced from Harper & Monks, 2023).

3.4 Implementation of framework

The model is built in Python 3.8 using the package SimPy (https://simpy.readthedocs.io/en/latest/), a process-based DES framework based on standard Python. We used Streamlit (https://docs.streamlit.io/), a Python library for building and deploying shareable data apps on the web. It provides a friendly, interactive front-end for the Python model, and among other methods, can be deployed free using Streamlit Community Cloud, albeit with resource limits. A further advantage to deploying a model in this way is that instructions for use and other documentation can be integrated into the app to support usability. All research artifacts are licensed using an MIT permissive license. All model code is accessible via GitHub (Orthopaedic Planning Model) and is archived and citable via Zenodo (Harper, 2023).
The orthopedic web app is available here: https://hospital-efficiency-project.streamlit.app/. It contains baseline values defined in the Python source code. It enables an operating theater schedule to be defined according to new rules (number of sessions per weekday, number of theaters in operation per weekday, surgical allocation per session). Other parameters can be changed in the web app: number of beds, mean lengths-of-stay per procedure, proportion of patients whose discharge is delayed, mean length-of-stay for delayed discharge. A user-defined schedule runs all scenarios with both the baseline schedule and the user-defined schedule and outputs results. The web app simulation can be accessed from any device.

For more technical users such as researchers or NHS analysts, the model is also available remotely via an executable scientific notebook using BinderHub. The code is interactive (e.g., baseline parameters can be changed, and other parameters can be explored in scenarios). Python is not required to be installed. Finally, documentation is provided using Jupyter Book and GitHub pages (Jupyter book HEP). This provides a structured website based on meta-data, markdown and Jupyter Notebooks (Executable Books Community, 2020), and includes scenario testing (automatically linked to BinderHub), and model documentation using STRESS guidelines (Monks et al. 2016).

4 RESULTS

4.1 Generating scenarios

![Web-app screenshot, annotated to highlight baseline and user-defined operating theater scheduling rules (blue), and scenarios table (yellow).](image)

Figure 3: Web-app screenshot, annotated to highlight baseline and user-defined operating theater scheduling rules (blue), and scenarios table (yellow).
The model is intended to provide additional decision-support, and not to provide definitive quantitative solutions. Its results are intended to be interpreted alongside other sources of information such as capital requirements, revenue, workforce availability and existing system knowledge.

Nonetheless, it provides an immediately usable source of information for business managers to support planning the requirements for the proposed hub, and for clinicians and service managers for planning its ongoing management, such as the effects of reducing lengths-of-stay in some procedures, or of eliminating all discharge delays through downstream capacity planning or alternative interventions.

Figure 3 (above) is an annotated screenshot of the web app illustrating an example where a new theater schedule has been generated as a user input to be used in combination with a set of simulation scenarios (Baseline scenario; Scenario 1 increased bed numbers; Scenario 2 increased bed numbers and reduced mean length-of-stay of patients with a discharge delay; Scenario 3 increased bed numbers, reduced mean length-of-stay of patients with a delayed discharge, and halved mean lengths-of-stay for all surgical procedures). Simulation scenarios are easily set up using sliders, and added to a displayed scenario table. The rules for the user-defined theatre schedule and the baseline theatre schedule (defined in Section 3.2) are displayed side-by-side for comparison and confirmation. The ‘allocations’ column allows users to flexibly allocate surgical classes into each of the chosen daily sessions (1R: one revision; 2P: two primaries; 1P: one primary; 2P_or_1R: a random allocation of two primaries or one revision). These allocations were defined by NHS staff. In the user-defined schedule, weekend activity is scheduled with two theaters in operation on Saturdays, and the same session allocations as baseline weekdays. If a user chooses to run the model with a user-defined schedule, the scenario table will display chosen simulation scenarios (highlighted in yellow) with both baseline and user-defined schedules (highlighted in blue).

4.2 Results visualization

High quality plots can be downloaded directly from the web application from each run of the model. An example set of results (Figures 4 and 5) show the effects on bed utilization and total surgical throughput (by procedure) of incrementally changing only the bed numbers, and using only the default theater schedule.

Figures 4 and 5: A range of bed numbers on (i) bed utilization; and (ii) total surgical throughput per week (by surgical type).
4.3 Model implementation

Given these and other results, it makes sense for business managers to work with clinicians to experiment with operating theater scheduling, lengths-of-stay, and delay scenarios aiming to reduce bed requirements. The model has been handed over in a series of handover events to the project team, orthopedic clinicians, business managers and service managers in the partner NHS Trust. Model outputs will provide supporting evidence for the Trust business case toward planning and resourcing the proposed high-volume low-complexity hub.

Additionally, we are promoting the model more widely with a view to testing it in further sites. This is timely, as many NHS regions are in the process of putting together business cases to access capital funds for elective planning against the new targets. In these cases, the underlying baseline parameters can be easily updated in the model source code. The model is also readily adaptable to other specialties.

5 CONCLUSION

5.1 Discussion and summary

This paper presents a free and open source discrete-event simulation (DES) web application for planning ring-fenced elective orthopedic activity for supporting elective COVID-19 wait list recovery. Our model was designed to address a real and specific capacity planning issue faced by the UK NHS, while maintaining the quality of reuse for similar applications. It therefore follows principles and guidelines defined in the conceptual modeling and generic modeling literature including keeping the model as simple as possible to facilitate understanding, to be flexible enough to support re-use, and to support incremental adaptation and modification where required (Robinson, 2008; Gunal, 2012; Penn et al. 2020). The model was intended to provide credible, rapid results for planning and implementation, so further advantages to simple modeling include model transparency, accessibility of assumptions, and ease of interpretation (Van der Zee, 2017; Tako, Tsioptias & Robinson, 2020).

In addressing a problem which is of immediate relevance to the NHS, the model also overcomes issues of perceived applicability and usefulness for decision-support (Brailsford et al. 2023; 2013). This includes the results themselves, the specific method (DES), but also as a means of conceptualizing healthcare processes in terms of stochastic servers, queues, and resources. While this was not a deliberate design decision, these broader concepts featured during the co-development process and are integral to interacting with the tool.

OR implementation research has found that to be implemented, results need to be accessible (Crowe et al. 2017; Monks, 2016). Usability is also a key feature (Brailsford et al. 2013; Harper, Mustafee & Pitt, 2022). Through application of our Open Modelling Sharing Framework (Harper & Monks, 2023), our simulation model provides readily implementable knowledge accessible in a usable and transparent format via a user-friendly, interactive web app. This removes the barrier of downloading and installing software and interacting with unfamiliar code or software to access and use results. Additionally, the model source code is openly available for re-use, and can be done without the requirement to install software. For users who wish to modify the source code, its licensing, dependency management, and documentation make this relatively straightforward.

5.2 Future work

The open approach used in this work means that as well as the availability of the web app for end-users in our partner Trust, the model is freely available for researchers and health-service planners to use, adapt or re-develop as required. As this is a topic of high current relevance, we are actively working with other healthcare services to investigate the applicability of the model for their own planning needs. The model is readily adaptable to other specialties, including the six prioritized within high-volume low-complexity planning to address high elective waits following the pandemic (GIRFT, 2023). There is also scope for
using this work as a pilot for wider high-volume low-complexity planning at regional or national levels across all specialties.

ACKNOWLEDGMENTS

This study was funded by the Health Data Research (HDR) UK Southwest Better Care Partnership (#6.12). The authors are supported by the National Institute for Health Research Applied Research Collaboration Southwest Peninsula. The views expressed in this publication are those of the author(s) and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care.

We wish to acknowledge our project partners: Rebecca Wilson, Maria Theresa Redaniel, Emily Eyles, Tim Jones, Chris Penfold, Andrew Elliott, Tim Keen, Ashley Blom and Andrew Judge, with additional thanks to Mike Whitehouse, Tom Woodward and Carolyn Roper.

REFERENCES


Harper, Pitt, and Monks


AUTHOR BIOGRAPHIES

ALISON HARPER is a Research Fellow at University of Exeter Peninsula Collaboration for Healthcare Operational Research and Data Science (PenCHORD). Her research interests include applied health and social care modeling and simulation, real-time simulation, and hybrid modelling. Her email address is a.l.harper@exeter.ac.uk.

MARTIN PITT is Professor of Applied Healthcare Modelling and Data Science at University of Exeter Medical School and Director of Peninsular Collaboration for Healthcare Operational Research and Data Science (PenCHORD). He is Director of UK Modelling and Simulation in Healthcare Network (MASHnet). His email address is m.pitt@exeter.ac.uk.

THOMAS MONKS is Associate Professor of Health Data Science at University of Exeter Medical School and Turing Fellow at the Alan Turing Institute. His research interests include open science for computer simulation, urgent and emergency care, and real-time discrete-event simulation. His website is https://arc-swp.nihr.ac.uk/about-penarc/people/tom-monks/ His email address is t.m.w.monks@exeter.ac.uk.