

This form documents the artifacts associated with the article (i.e., the data and code supporting the computational findings) and describes how to reproduce the findings.

Part 1: Data

- ☐ This paper does not involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).
- ☒ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Abstract

This SEER breast cancer dataset was obtained from the 2017 November update of the SEER Program of the NCI, which provides information on population-based cancer statistics. The dataset involves female patients with infiltrating duct and lobular carcinoma breast cancer (SEER primary cites recode NOS histology codes 8522/3) diagnosed in 2006-2010. The observation on the survival time has been formatted into the needed interval-censored data. The patients with unknown tumor size, examined regional LNs, regional positive LNs, and the patients whose survival months were less than 1 month were excluded; thus 4024 patients were ultimately included.

Availability

- ☒ Data **are** publicly available.
- ☐ Data **cannot be made** publicly available.

If the data are publicly available, see the *Publicly available data* section. Otherwise, see the *Non-publicly available data* section, below.

Publicly available data

- ☐ Data are available online at:
- ☐ Data are available as part of the paper's supplementary material.
- ☒ Data are publicly available by request, following the process described here:

The dataset is available through an Internet connection or a DVD. To access the SEER data, interested users are required to sign a SEER Research Data Agreement form (available at <https://seer.cancer.gov/data/sample-dua.html>) and then submit a formal request for access to the data. More details about where to access to the SEER data can be found on the website <https://seer.cancer.gov/data/access.html>.

- ☐ Data are or will be made available through some other mechanism, described here:

Non-publicly available data

Description

File format(s)

- ☐ CSV or other plain text.
- ☒ Software-specific binary format (.Rda, Python pickle, etc.): .mat
- ☐ Standardized binary format (e.g., netCDF, HDF5, etc.):
- ☐ Other (please specify):

Data dictionary

- ☒ Provided by authors in the following file(s): README.pdf
- ☐ Data file(s) is(are) self-describing (e.g., netCDF files)
- ☐ Available at the following URL:

Additional Information (optional)

Part 2: Code

Abstract

This code provides a MATLAB implementation for the estimation and variable selection of interval-censored failure time data with a random change point, along with an application to the SEER breast cancer study. The MATLAB scripts are designed to replicate the analyses, plots, and tables presented in the paper.

Description

Code format(s)

- ☒ Script files
 - ☐ R
 - ☐ Python
 - ☒ Matlab
 - ☐ Other:
- ☐ Package
 - ☐ R
 - ☐ Python
 - ☐ MATLAB toolbox
 - ☐ Other:
- ☐ Reproducible report
 - ☐ R Markdown
 - ☐ Jupyter notebook
 - ☐ Other:
- ☐ Shell script
- ☐ Other (please specify):

Supporting software requirements

Version of primary software used MATLAB R2017b; MATLAB R2022b

Libraries and dependencies used by the code Not Applicable

Supporting system/hardware requirements (optional)

Parallelization used

- ☐ No parallel code used
- ☒ Multi-core parallelization on a single machine/node
 - Number of cores used: 12
- ☐ Multi-machine/multi-node parallelization
 - Number of nodes and cores used:

License

- ☒ MIT License (default)
- ☐ BSD

- ☐ GPL v3.0
- ☐ Creative Commons
- ☐ Other: (please specify)

Additional information (optional)

Part 3: Reproducibility workflow

Scope

The provided workflow reproduces:

- ☒ Any numbers provided in text in the paper
- ☒ The computational method(s) presented in the paper (i.e., code is provided that implements the method(s))
- ☒ All tables and figures in the paper
- ☐ Selected tables and figures in the paper, as explained and justified below:

Workflow

Location

The workflow is available:

- ☒ As part of the paper's supplementary material.
- ☐ In this Git repository:
- ☐ Other (please specify):

Format(s)

- ☐ Single master code file
- ☐ Wrapper (shell) script(s)
- ☐ Self-contained R Markdown file, Jupyter notebook, or other literate programming approach
- ☒ Text file (e.g., a readme-style file) that documents workflow
- ☐ Makefile
- ☐ Other (more detail in *Instructions* below)

Instructions

The main code is located in the `IC_RandomChangePoint` folder, which includes the `IC_Functions` subdirectory and the primary functions needed to replicate the simulation tables and figures in Section 5 as well as the SEER breast cancer data analysis in Section 6 of the paper "Estimation and Variable Selection for Interval-censored Failure Time Data with Random Change Point and Application to Breast Cancer Study". For different specified transformation models, adjust the `Gx.m` function accordingly. For further details and reproduction instructions, please refer to the `README` file.

Expected run-time

Approximate time needed to reproduce all analyses on a standard desktop machine:

- ☐ < 1 minute
- ☐ 1-10 minutes
- ☐ 10-60 minutes
- ☒ 1-8 hours
- ☐ > 8 hours
- ☐ Not feasible to run on a desktop machine, as described here:

Additional information (optional)

Notes (optional)