Author Contributions Checklist

Alexander Wolfgang Jung

May 03, 2021

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

Publicly Available:

The data that is publicly availbe was taken from R-Survival (Therneau and Lumley 2015). Colon are data on trials of adjuvant chemotherapy for colon cancer, (Laurie et al. 1989). Lung are data extracted from the North Central Cancer Treatment Group on mortality for advanced lung cancer (Loprinzi et al. 1994). NAFLD is a large population-based study investigating non-alcoholic fatty liver disease (NAFLD) (Allen et al. 2018). Heart investigates mortality in patients from the Stanford heart transplant program (Crowley and Hu 1977)

Non-publicly available:

The UK Biobank (UKB) (Sudlow et al. 2015) is a large-scale biomedical database established in 2006. The information collected on individuals includes questionnaires and face-to-face interviews (covering general medical factors, lifestyle, environmental factors, socioeconomics, etc.), physical measurements, blood and urine assays, prescriptions, and genotypes, as well as linkage to their hospital admission records. Further, detailed information, like wearable devices, precise dietary information, is available for smaller subsets. In total there are 502628 participants, recruited between 2006 and 2010. All participants were between 40-69 years of age at their recruitment date. We study the association of standard risk factors and comorbidities, taken from the electronic health records, with the occurrence of a myocardial infarction. For details of the analysis see section 4 of the paper.

Availability

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

Publicly available data

☒ Data are available online at: https://github.com/alexwjung/ProbCox

☒ Data are available as part of the paper's supplementary material.

☐ Data are publicly available by request, following the process described here:

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

Non-publicly available data

The UK-Biobank (UKB) data contains sensible information on individuals and cannot be openly shared for privacy considerations. However, researchers can apply for access to the UKB under: https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access

The simulation study we present in the paper closely resembles the data structure encountered in electronic health records, similar to the data in the UKB, and can be used as a guidance.

Electronic health records and biobanks contain a vast array of information on individuals that open new possibilities to study disease and are therefore highly important in biomedical research. However, the wealth of information also makes it difficult to obfuscate data in such a way as to preserve important structures in the data. This inevitably leads to issue for openly distributing the data.

Description

The data provided can be found in Data

File format(s)

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

Data dictionary

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

Additional Information (optional)

Part 2: Code

Abstract

We provide a python package (probcox) for the model implemented. This is a scalable version of the standard Cox model fitted via stochastic variational inference. The code is written entirely in pytorch/pyro and numpy. We provide an algorithm to efficiently simulate survival times with time-varying covariates that resemble disease association studies in electronic health records. The original scripts for all the analyses are provided.

Additionally, there are jupyter notebooks that can be run in google colab to readily replicate all of the results presented (except for the UKB as we cannot provide the data as mentioned above).

Description

Code format(s)

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

Supporting software requirements

We provide replicable notebooks that can be used with Google Colab (requirement: Google account). The notebooks can also be run via standard jupyter notebooks.

Version of primary software used Python version 3.7

R version 4.0.3

Libraries and dependencies used by the code We use conda [https://docs.conda.io/en/latest/] as a package manager and the full list of install packages can be found/replicated via R environment and Python environment, respectively.

Supporting system/hardware requirements (optional)

We had access to a cluster to run the simulations. The main purpose for this was to run simulation runs independently.

Parallelization used

- ☐ No parallel code used
- ☒ Multi-core parallelization on a single machine/node
  - Number of cores used: varying - can be adjusted by user - will be clear in scripts.
- ☐ Multi-machine/multi-node parallelization
  - Number of nodes and cores used:

License

☒ MIT License (default)
☐ BSD
☐ GPL v3.0
☐ Creative Commons
☐ Other: (please specify below)

Additional information (optional)

The code for the model can be installed via:

```
$ pip install probcox
```

Scope

The provided workflow reproduces:

☒ Any numbers provided in text in the paper
☒ All tables and figures in the paper
☐ Selected tables and figures in the paper, as explained and justified below:

Workflow

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

Expected run-time

Approximate time needed to reproduce the 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)

- Rerunning all the simulations on a single desktop machine will take a considered amount of time. We therefore provide individual simualtion runs (choosen by demand) that can be checked/compared to the results provided on https://github.com/alexwjung/ProbCox.

Notes (optional)

The code for the UKB analysis as well as all the scripts for data preperations are also provided UKB.

References

Allen, Alina M, Terry M Therneau, Joseph J Larson, Alexandra Coward, Virend K Somers, and Patrick S Kamath. 2018. "Nonalcoholic Fatty Liver Disease Incidence and Impact on Metabolic Burden and Death: A 20 Year-Community Study." *Hepatology* 67 (5): 1726–36.

Crowley, John, and Marie Hu. 1977. "Covariance Analysis of Heart Transplant Survival Data." *Journal of the American Statistical Association* 72 (357): 27–36.

Laurie, John A, Charles G Moertel, Thomas R Fleming, Harry S Wieand, John E Leigh, Jebal Rubin, Greg W McCormack, James B Gerstner, James E Krook, and James Malliard. 1989. "Surgical Adjuvant Therapy of Large-Bowel Carcinoma: An Evaluation of Levamisole and the Combination of Levamisole and Fluorouracil. The North Central Cancer Treatment Group and the Mayo Clinic." *Journal of Clinical Oncology* 7 (10): 1447–56.

Loprinzi, Charles Lawrence, John A Laurie, H Sam Wieand, James E Krook, Paul J Novotny, John W Kugler, Joan Bartel, Marlys Law, Marilyn Bateman, and Nancy E Klatt. 1994. "Prospective Evaluation of Prognostic Variables from Patient-Completed Questionnaires. North Central Cancer Treatment Group." *Journal of Clinical Oncology* 12 (3): 601–7.

Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age." *Plos Med* 12 (3): e1001779.

Therneau, Terry M, and Thomas Lumley. 2015. "Package 'Survival'." *R Top Doc* 128 (10): 28–33.