Author Contributions Checklist Form

Commented [A1]: This document will be read by readers after an accepted manuscript is published so please phrase your responses for future readers. Please see https://jasa-acs.github.io/repro-guide/pages/acc.html for details, links to examples, and information on how to save your completed form as a PDF. In particular make sure to remove these comments when saving as a PDF.

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).

Commented [A2]: If this box is checked, please skip directly to the Code section. Otherwise, continue.

☒ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Abstract

Commented [A3]: A short (< 100 words) description of the data. More details can be provided in files accompanying the data.

Our data are collected in a study called HOPE-B. HOPE-B is a phase 3 study of the first FDA-approved gene therapy for Hemophilia B in adults (NCT03569891). The study enrolled 54 men, and as of the time of our data analysis, data from up to 18 months post-treatment were available, with ongoing data collection for up to 5 years. Our goal is to make predictions on the efficacy of the gene therapy based on factor IX activity levels. To improve predictions, we consider the external information from a phase 1 study (NCT02396342) and a phase 2b study (NCT03489291).

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:

Commented [CP4]: If the data are only available by request to the authors or some other data owner, please make sure to explain the process of requesting access to the data.

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

Non-publicly available data

Commented [A5]: If the data for this manuscript are publicly available, skip to the Description section below. Otherwise, continue.

The Journal of the American Statistical Association requires authors to make data accompanying their papers available to the scientific community except in cases where: 1) public sharing of data would be impossible, 2) suitable synthetic data are provided which allow the main analyses to be replicated (recognizing that results may differ from the "real" data analyses), and 3) the scientific value of the results and methods outweigh the lack of reproducibility.

Discussion of lack of publicly available data:

Commented [A6]: For example:

- why data sharing is not possible,
- what synthetic data are provided, and
- why the value of the paper's scientific contribution outweighs the lack of reproducibility.

The data are collected from an ongoing clinical trial, which contains patients' confidential information. For privacy, we have to keep the data non-shared.

We need some discussions on the third point. Say the predictions help provide evidence of the efficacy of the gene therapy.

Description

File format(s)

Commented [CP7]: Check all that apply.

☐ CSV or other plain text:

☐ Software-specific binary format (.Rda, Python pickle, etc.):

☐ Standardized binary format (e.g., netCDF, HDF5, etc.):

☐ Other (described here):

Data dictionary

Commented [CP8]: A data dictionary provides information that allows users to understand the meaning, format, and use of the data.

☐ Provided by the authors in the following file(s):

☐ Data file(s) is (are) self-describiing (e.g., netCDF files)

☐ Available at the following URL:

Additional information (optional)

Commented [CP9]: Provide any details that would be helpful in understanding the data. If relevant, provide unique identifier/DOI/version information and/or license/terms of use.

Journal of the American Statistical Association

Version: 2023-10-11

Part 2: Code

Abstract

Commented [A10]: A short (< 100 words) description of the code. If necessary, more details can be provided in files that accompany the code. If no code is provided, please state this and say why (e.g., if the paper contains no computational work).

The code for real data analysis and simulation is compiled in Rcpp with built-in functions in RcppArmadillo for faster implementation. The code realizes sampling from a Bayesian averaging model, with the income being internal and external trajectories, arranged in R lists, and the outcome representing the posterior samples from our proposed model.

Description

Code format(s)

Commented [CP11]: Check all that apply.

☒ Script files

☒ R ☐ Python ☐ Matlab

☐ Other:

☒ Package

☒ R ☐ Python ☐ MATLAB toolbox

☐ Other:

☒ Reproducible report

☒ R Markdown ☐ Jupyter notebook

☐ Other:

☐ Shell script

☐ Other (described here):

Supporting software requirements

Commented [A12]: Please cite all software packages in the References Section in similar fashion to paper citations, citing packages that are foundational to the research outcome (including packages that implement methods to which you compare your methods). You may elect to not cite packages used for supporting purposes. For R packages, note that running `citation('name_of_package')` often shows how the package authors wish to be cited.

Version of primary software used

Commented [CP13]: For example, R version 3.6.2.

Local: R version 4.0.2.
Server: R version 4.0.4.

Libraries and dependencies used by the code

Commented [CP14]: Include version numbers (e.g., version numbers for any R or Python packages used)

Rcpp: 1.0.5

Dirk Eddelbuettel and Romain Francois (2011). Rcpp: Seamless R and C++ Integration. Journal of
Statistical Software, 40(8), 1-18. URL http://www.jstatsoft.org/v40/i08/.

Spatstat: 1.64.1

Adrian Baddeley, Ege Rubak, Rolf Turner (2015). Spatial Point Patterns: Methodology and Applications with
R. London: Chapman and Hall/CRC Press, 2015. URL
http://www.crcpress.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/

rlang: 1.1.1

Lionel Henry and Hadley Wickham (2023). rlang: Functions for Base Types and Core R and 'Tidyverse'
Features. R package version 1.1.1. https://CRAN.R-project.org/package=rlang

permute: 0.9.5

Gavin L. Simpson (2019). permute: Functions for Generating Restricted Permutations of Data. R package
version 0.9-5. https://CRAN.R-project.org/package=permute

Supporting system/hardware requirements (optional)

Commented [A15]: System/hardware requirements including operating system with version number, access to cluster, GPUs, etc.

The code requires 100 cores with each running a simulation task. For real data analysis, 100 cores are required to conduct 100 replications with different initializations.

Parallelization used

☐ No parallel code used
☒ Multi-core parallelization on a single machine/node
Number of cores used: 100
☐ Multi-machine/multi-node parallelization
Number of nodes and cores used:

License

☐ MIT License (default)
☐ BSD
☐ GPL v3.0
☐ Creative Commons
☐ Other (described here):

Additional information (optional)

Commented [CP16]: By default, submitted code will be published on the JASA GitHub repository (https://github.com/jasa-acs) as well as in the supplementary material. Authors are encouraged to also make their code available in a public code repository, such as GitHub, GitLab, or BitBucket. If relevant, please provide unique identifier/DOI/version information (e.g., a Git commit ID, branch release or tag). If the code and workflow are provided together, this section may be omitted, with information provided in the "Location" section below.

Part 3: Reproducibility workflow

Commented [CP17]: The materials provided should provide a straightforward way for reviewers and readers to reproduce analyses with as few steps as possible.

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 here:

The following results in Table 1 will be reproduced:
DGP1, K = 20, rho = 0
DGP4, K = 20, rho = 0.5

Workflow details

Location

Commented [CP18]: Check all that apply, and in the case of a Git repository include unique identifier, such as specific commit ID, branch, release, or tag.

Indicate where the materials (generally including the code, unless in a separate location and indicated in the previous section) are available. We strongly encourage authors to place their materials (but not large datasets) in a Git repository hosted on a site such as GitHub, GitLab, or BitBucket. If the repository is private during the review process, please indicate the location where it will be available publicly upon publication, and also include the materials as a zip file (e.g., obtained directly from the Git hosting site) as supplementary materials.

The workflow is available:
☐ As part of the paper's supplementary material
☒ In this Git repository:
☐ Other:

Format(s)

Commented [CP19]: Check all that apply

☐ 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

Commented [A20]: Describe how to use the materials provided to reproduce analyses in the manuscript. Additional details can be provided in file(s) accompanying the reproducibility materials. If no workflow is provided, please state this and say why (e.g., if the paper contains no computational work).

The provided .md file only demonstrates the result given seed 1. To reproduce the Table 1 result, one needs to run the file for seeds 1 to 100. To get the results using parallel computing, the instruction is in Section 3. The summary code is included in the chunks in Section 2 of the .md file.

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:

For each simulation task, it takes around 8 hours to 16 hours to finish. One needs to consider parallel

Additional documentation (optional)

Commented [A21]: Additional documentation provided (e.g., R package vignettes, demos or other examples) that show how to use the provided code/software in other settings.

Notes (optional)

Commented [A22]: Any other relevant information not covered on this form. If reproducibility materials are not publicly available at the time of submission, please provide information here on how the reviewers can view the materials (and make sure to remove this information when submitting the final version of this form for an accepted manuscript).