Issues in the planning and reporting of studies that assess performance of statistical & computational methods

with emphasis on high-dimensional data

LISA MCSHANE

on behalf of TG9 (High-dimensional Data Topic Group), and based heavily on presentations and published letter by Simulation Panel members A. Boulesteix, A. Benner, H. Binder, M Abrahamowicz, and W. Sauerbrei

Need for method performance assessment
(Boulesteix et al., Biometrical Journal 2018;60:216-218 [Letter])

- For many areas of statistical application there are already a large number of methods available, but far less guidance on which methods are optimal or even appropriate for particular situations
- Chances of publication in a statistics or computational journal are much higher when a “new” method is being proposed, but performance assessments may be limited and/or biased
- Many new methods are complex and properties are often not possible to assess based on theoretical arguments, or may require strong and possibly unrealistic assumptions

Two main approaches to performance assessment

- Demonstrate method on “real” data
  - Challenging to find multiple data sets for which method is applicable
  - Might not know “truth” unless data were generated from a controlled experiment
- **Simulation studies**
  - Imperfect reflection of reality
  - “Reality” may be too complex to adequately capture through usual purely model-based simulations (especially for high-dimensional data)

Risk of bias in published performance assessments

- New method developed to address features of a particular data set, and performance addressed only on that data set
- New method evaluated on multiple data sets; results reported only for data sets on which the new method performed best
- Simulations engineered to generate data with features that the new method is designed to leverage
  - Example: Pooling or "borrowing information" over parameter estimates or subsets
- New method developers have greater expertise in applying their own methods; possibly no involvement of "advocate/expert" for competing method

Key steps and decisions in the planning , coding, analysis, and reporting of simulation studies

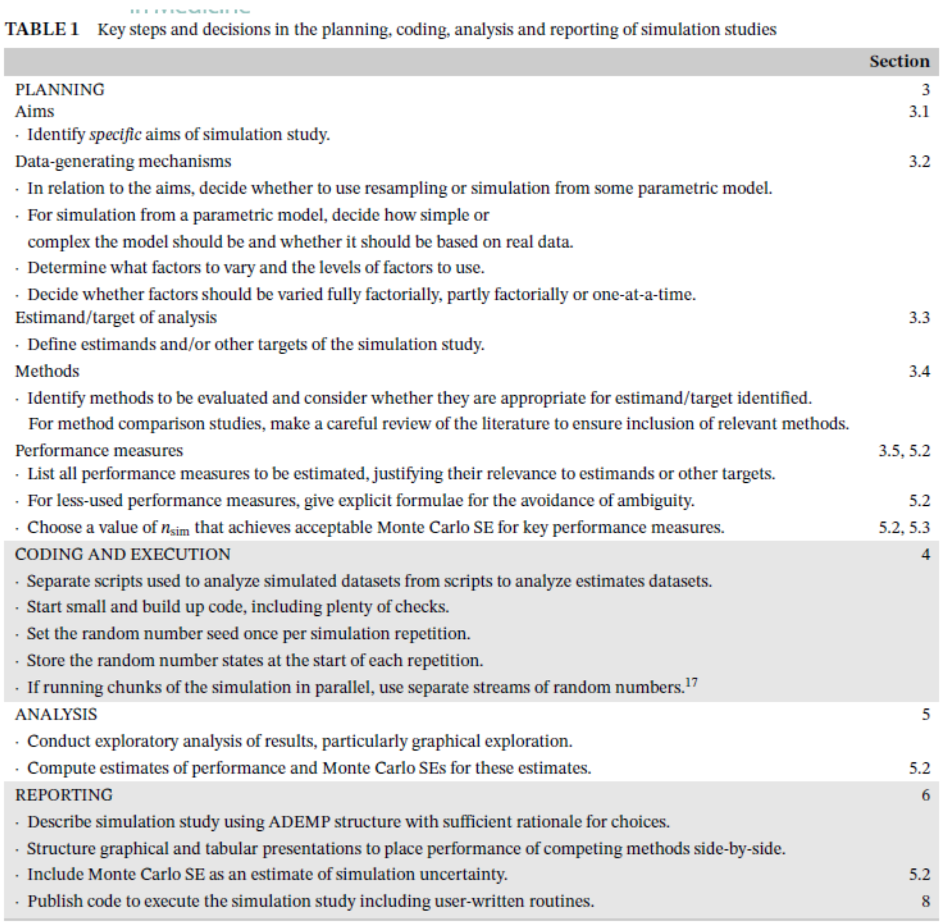

TABLE 1 Key steps and decisions in the planning, coding, analysis and reporting of simulation studies

	Section
PLANNING	3
Aims	3.1
· Identify *specific* aims of simulation study.	
Data-generating mechanisms	3.2
· In relation to the aims, decide whether to use resampling or simulation from some parametric model.	
· For simulation from a parametric model, decide how simple or complex the model should be and whether it should be based on real data.	
· Determine what factors to vary and the levels of factors to use.	
· Decide whether factors should be varied fully factorially, partly factorially or one-at-a-time.	
Estimand/target of analysis	3.3
· Define estimands and/or other targets of the simulation study.	
Methods	3.4
· Identify methods to be evaluated and consider whether they are appropriate for estimand/target identified. For method comparison studies, make a careful review of the literature to ensure inclusion of relevant methods.	
Performance measures	3.5, 5.2
· List all performance measures to be estimated, justifying their relevance to estimands or other targets.	
· For less-used performance measures, give explicit formulae for the avoidance of ambiguity.	5.2
· Choose a value of n_{sim} that achieves acceptable Monte Carlo SE for key performance measures.	5.2, 5.3
CODING AND EXECUTION	4
· Separate scripts used to analyze simulated datasets from scripts to analyze estimates datasets.	
· Start small and build up code, including plenty of checks.	
· Set the random number seed once per simulation repetition.	
· Store the random number states at the start of each repetition.	
· If running chunks of the simulation in parallel, use separate streams of random numbers.[17]	
ANALYSIS	5
· Conduct exploratory analysis of results, particularly graphical exploration.	
· Compute estimates of performance and Monte Carlo SEs for these estimates.	5.2
REPORTING	6
· Describe simulation study using ADEMP structure with sufficient rationale for choices.	
· Structure graphical and tabular presentations to place performance of competing methods side-by-side.	
· Include Monte Carlo SE as an estimate of simulation uncertainty.	5.2
· Publish code to execute the simulation study including user-written routines.	8

Morris et al., *Statistics in Medicine* 2019;38:2074–2102.

Structured approach for planning and reporting simulation studies (“**ADEMP**”)

- **A**ims *of the simulation study*
- **D**ata-generating mechanisms
- **E**stimands *or other targets of the simulation study*
- **M**ethods *to be evaluated*
- **P**erformance measures

Special considerations for simulation studies involving high-dimensional data (HDD)

- Aims, estimands, and performance metrics may be complex

  Examples

  - Which method produces a classifier/predictor that ***performs best***?
    - Recall yesterday's discussion of model/predictor performance assessment
  - Which method most accurately identifies the true ***clusters***?
    - Can we even define the notion of a cluster?
  - Which method most accurately identifies ***gene networks***?
    - Airport discussion with Mitch Gail

Special considerations for simulation studies involving HDD (cont.)

- Methods to be evaluated may be complex, multi-step processes involving sophisticated algorithms
  - Access to computer code may be required to implement the methods
    - Coding languages may be different (e.g., R, STATA, MatLab, Python)
  - Successful implementation of method may require substantial expertise
    - Options, tuning parameters, convergence, etc.
  - Access to high performance computing facility

Special considerations for simulation of HDD
(next several slides borrow from lecture of A. Benner 3/21/18)

- Fundamental difficulties in simulating HDD
  - Simulation of completely synthetic data cannot capture complex correlation structure among covariates in HDD
  - Underlying mechanism (e.g., biological) not well understood
    - Difficult to propose suitable multivariable model relating HDD (e.g., molecular) and/or covariates to dependent variable
  - Some characteristics of HDD are not uniquely defined (e.g., “cluster”)
- Investigation of asymptotic behavior may require **EXTREMELY LARGE** n!

Special considerations for simulation of HDD (cont.)

- Completely parametric data generating mechanisms challenging to implement
  - Simulations based on assumed distributions (e.g., multivariate Gaussian, Poisson or negative binomial for count data such as from RNAseq)
    - How to simulate correlated non-Gaussian data?
    - What are realistic effects and correlation structures?
  - Simulations based on a model with parameters estimated from pilot data
    - Imprecise estimates of parameters (e.g., number of parameters in variance-covariance matrix is more than # of observations when p>>n)

“Real data” simulation of HDD

Useful approach for realistic HDD generation

- Plasmode data: Real data (e.g., omics data from actual biological specimens) which are manipulated such that the parameters of interest are known with certainty.
  - Name from plasm=form, and mode=measure
  - References:
    - Cattell, R. B. (1966). Handbook of Multivariate Experimental Psychology. Rand McNally psychology series. Rand McNally, Chicago.
    - Mehta et al., Physiological Genomics 2006;28(1):24-32

“Real data” simulation of HDD

- Advantages of plasmode data
  - Distributions/correlations are taken directly from real data
  - Appropriate permutation, resampling, or modification of real data offers flexibility to generate data with desired features
  - Can combine with outcome models to generate dependent variables associated with realistic HDD as independent variables

“Real data” simulation of HDD

More on plasmode-type approaches

Example 1: Generate data for evaluation of multiple testing methods

- Permute subject/specimen IDs to generate a null distribution
  - Global null allows assessment of “weak control” of false positives for a multiple testing procedure
- Add back defined effects on specific individual variables
  - Allows assessment of both “power” for true positives and “strong control” of false positives for a multiple testing procedure

“Real data” simulation of HDD

More on plasmode-type approaches

Example 2: Generate mixture distributions

- Mix distinct data sets in varied proportions, e.g., mixture of molecular profiles of two or more species of gut bacteria
  - Mitch Gail airport discussion

“Real data” simulation of HDD

More on plasmode-type approaches

Example 3: Generate clustered data

- Merge HDD from classes with distinct (high-dimensional) means and add noise or dilate mean distances to generate data sets with less or more separated clusters, respectively
  - Jörg Rahnenführer talk at a statistical meeting in early 2000s

“Real data” simulation of HDD

More on plasmode-type approaches

<u>Example 4</u>: HDD data as the dependent variables

$$X_i = g(age, gender, \ldots), \quad j = 1, 2, \ldots, p$$

<u>Example 5</u>: HDD as the explanatory variables

$$Y = h(X_1, X_2, \ldots, X_p, age, gender, \ldots)$$

“Real data” simulation of HDD

More on plasmode-type approaches

Example 6: Generate cohort data with HDD confounding

- Sample with replacement from cohort data to get desired samples size *n* and event rate
- Calculate $p_i = P(Y_i = 1 | E_i, \boldsymbol{X}_{ic})$, $i = 1, 2, ..., n$, for desired model where E_i = exposure, $\boldsymbol{X}_{ic}$ = HDD vector of confounders.
- Simulate binary outcome status according to

$$Y_i^* \sim Binomial(1, p_i),\ i = 1, 2, ..., n$$

Summary remarks

- Great need for assessment of performance of HDD methods
- Number of “real” HDD sets available will always be too small relative to the multitude of data types, cohort characteristics, analytical goals and methods
- STRATOS could provide a great service by educating on valid and useful approaches for simulation studies involving HDD
- DISCUSSION?