Issues in Modern Biomedical Simulation Studies
with emphasis on "omics" studies

Victor Kipnis
US National Cancer Institute

on behalf of the STRATOS simulation panel

Need for performance assessment

(Boulesteix et al., Biometrical Journal, 2018)

- In statistical applications, there are many available methods but little guidance on their performance or their comparison for particular situations
- Chances of publication increase when a “new” method is proposed, but assessment of its performance may be limited
- Statistical properties of complex methods are unknown or may require strong and possibly unrealistic assumptions

Need for performance assessment

- With frequent lack of a clear biological rationale, omics high-dimensional data are particularly prone to overfitting
- Statistical model selected from the discovery phase may function well on the samples used for the discovery research but is inaccurate on any other sample
- Moreover, the process of selection makes analytical derivation of the statistical properties of the selected model all but impossible

Main approaches to performance assessment

- Use of "real" independent data sets
  - challenging to find multiple data sets of the same quality and reflecting the same phenomenon, especially in omics studies
  - in some cases "truth" is unknown which allows only to assess reproducibility of (often biased) results but not their properties

Simulation studies

- Simulation studies are computer experiments that involve creating data with known true structure in pre-defined scenarios
  - may involve imperfect reflection of reality
  - ability to evaluate appropriateness and performance of current and novel statistical methods in considered scenarios
- Simulation studies become especially critical in current omics research that generates complex high-dimensional data with limited sample size

Risk of bias in performance assessment in simulations

- New method is developed to address a particular data set, and its performance addressed only on that data set
- New method is evaluated on multiple data sets but results are reported only for data sets on which the new method performed best
- Simulations are engineered to generate data with features that the new method is designed to leverage
- New method developers have greater expertise in applying their own methods and use "tailored" simulations

Neutral comparison studies

- In contrast, neutral comparison studies are dedicated to assessment/comparison itself:
  - they do not aim to demonstrate the superiority of a particular method and are not designed in a way that may increase the probability to observe incorrectly this superiority
  - they involve authors who are approximately equally competent on all considered methods
- Thus neutral comparison studies can be considered as unbiased

Structured approach for planning and reporting simulation studies ("ADEMP")

(Morris et al., Statistics in Medicine 2019)

- Aims of the simulation study
- Data-generating mechanisms
- Estimands or other targets of the simulation study
- Methods to be evaluated
- Performance measures

Special consideration for simulation of omics studies

- Aims, estimands, and performance metrics may be complex
  - which method produces the best predictor/classifier?
  - which method more accurately identifies true clusters?
  - which method more accurately identifies gene network?

Special consideration for simulation of omics studies

- Data generating requires careful consideration:
  - completely parametric data generating is challenging to implement due to unknown probability distributions, realistic effects, and correlational structure
  - resampling from a real data set will produce biased results if methods involve selection of "best" features (resampling, e.g., bootstrap, works only for evaluating smooth procedures)

Special consideration for simulation of omics studies

- Data generating requires careful consideration:
  - useful approach may involve plasmode data (name from plasm = form, and mode = measure)
  - e.g., using real data to generate distribution of multiple covariates, while generating outcomes is based on assumed regression form/parameters

Special consideration for simulation of omics studies

- Important issue: how to generate data for comparing different procedures
  - should generated data sets be independent not only for assessing each method but across methods as well?
  - should different procedures be compared using same data sets (similar to paired t-test)?

Special consideration for simulation of omics studies

- Assessment of model selection procedures should be based NOT only on assessment of a particular model selected by it using training data set but the procedure itself (similar to the difference between estimator and a particular estimate that it produces)
  - e.g., variance of AUC on validation data D^v for a particular classifier P selected on training data D^{tr} is

$$E[var\{\text{AUC}^p\left(D^{tr}, D^v\right)|D^{tr}\}] + var[E\{\text{AUC}^p\left(D^{tr}, D^v\right)|D^{tr}\}]$$

Additional issues with simulation studies

- What parameters and assumptions should be varied across the simulated scenarios?
- What range of sample sizes should be assessed?
- How can we assess the practical relevance of simulation results, which depends on the real-life plausibility of the simulation scenarios?
- How can an acceptable neutrality of the authors team be achieved and how can non-neutrality (the analog of "conflicts of interest" in clinical research) be disclosed?

Conclusions

- To improve assessment of statistical methods and their reproducibility it is desirable to
  - (i) reinforce the status of neutral comparison studies
  - (ii) develop large collaborative research on how to reliably assess statistical methods
  - (iii) derive reporting guidelines to increase transparency of assessment/comparison studies, similar to existing guidelines for many types of studies in the health sciences