Can and should the choice of statistical methods be more "evidence-based"?

Anne-Laure Boulesteix
joint with Alexander Hapfelmeier

Institute for Medical Informatics, Biometry and Epidemiology
Ludwig-Maximilians-University Munich

Warwick, September 16th 2016
CRiSM Workshop: Contemporary Issues in Hypothesis Testing

Benchmarking experiments with real datasets

- compute cross-validation error of K prediction methods for J data sets , e.g. from repositories
- test difference in prediction error using paired t-test or Wilcoxon test to compare methods 1 and 2

	e_1	e_2
D_1	$e_1(D_1)$	$e_2(D_1)$
$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$
D_J	$e_1(D_J)$	$e_2(D_J)$

Often: $J \approx$5–15, $K \approx$2–10 (e.g., kNN, linear discriminant analysis, random forest, support vector machines, etc.)

Analogy between clinical trials and benchmarking

datasets $D_1, \ldots, D_J$	$\approx$	patients
methods $M_1, \ldots, M_K$	$\approx$	therapies
computational scientist	$\approx$	trialist
applied data analyst	$\approx$	medical doctor
prediction error	$\approx$	primary endpoint
datasets' characteristics $(n, p, \text{etc.})$	$\approx$	biomarkers

Evidence-based medicine (EBM)

Greenhalgh (British Medical Journal, 2014) states:

"It is more than 20 years since evidence-based medicine working group announced a "new paradigm" for teaching and practicing clinical medicine. Tradition, anecdote, and theoretical reasoning from basic sciences would be replaced by evidence from high quality randomized controlled trials and observational studies, in combination with clinical expertise and the needs and wishes of patients."

$\rightarrow$ Randomized clinical trials play a central role towards evidence-based medicine.

In this talk

What about an "evidence-based" data analysis in which "tradition, anecdote, and theoretical reasoning from basic sciences [including simulations] would be [complemented] by evidence from high-quality [benchmark studies], in combination with [statistical] expertise and the needs and wishes of the [substantive scientists]"?

- not all principles from EBM can be transferred to benchmark experiments in statistical research;
- but some of them are helpful to make statistical research more "evidence-based";
- our considerations are meant as feed for thoughts on scientific practice, not as guidelines or criticism.

Randomized clinical trials

History:

- (Almost) randomized trial by James Lind (1747): treatment of scurvy
- 2 patients allocated to each of cider, elixir vitriol, vinegar, nutmeg, sea water, and oranges/lemons. Those given oranges and lemons showed the "most sudden and visible good effects".
- well-established methodology for 50 years

Key principles:

- random assignment of the patients to one of the therapy groups, e.g. standard therapy vs. new intervention
- single or double blinding
- strict inclusion criteria
- precise analysis plans and protocols
- sample size calculation

Randomized clinical trials methodology: testing

- For a normally distributed endpoint, the *treatment effect* is the difference between the means of the endpoint in the two groups.
- The t-test is used to test the null-hypothesis that this difference is zero.
- A point estimator and a confidence interval for this difference can simply be obtained.
- The appropriate sample size to detect a difference considered clinically different given the assumed within-group variance of the endpoint with a power of 80% at a significance level of 0.05 is determined before starting the trial.

Randomized clinical trials methodology: which patients to include in the analysis?

- Strict inclusion criteria are defined, e.g. "age>18", "no pregnancy/breastfeeding", unilateral disease, no diabetes, etc.
- All patients satisfying the criteria and giving their informed consent are enrolled in the trial.
- Patients who drop-out from the study arm they were included in are given much attention.
- Per-protocol or intention-to-treat analyses are conceivable.
- Subgroup analyses as defined prior to data collection are considered relevant.

PLOS | COMPUTATIONAL BIOLOGY

EDITORIAL

Ten Simple Rules for Reducing Overoptimistic Reporting in Methodological Computational Research

Anne-Laure Boulesteix*

Institute for Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians University, Munich, Germany

	e_1	e_2
D_1	$e_1(D_1)$	$e_2(D_1)$
...	...	...
...	...	...
...	...	...
...	...	...
...	...	...
D_J	$e_1(D_J)$	$e_2(D_J)$

- Rule 1: Assess the New Method
- Rule 2: Compare the New Method to the Best
- Rule 3: **Consider Enough Datasets**
- Rule 4: **Do Not "Fish" for Datasets**
- Rule 5: **Think of the No-Free-Lunch Theorem and Report Limitations**
- Rule 6: **Consider Several Criteria**
- Rule 7: Validate Using Independent Data
- Rule 8: Design Simulations Appropriately
- Rule 9: Provide All Information
- Rule 10: Read the Other Ten Simple Rules Articles

What is being tested?

- Distribution P of data is considered as the outcome of a random variable Φ, and size of data set n as the outcome of a random variable N.
- Then the hypothesis that is implicitly being tested when comparing methods M_1 and M_2 can be written as

$$\boldsymbol{E}(\varepsilon(M_1, \Phi, N)) = \boldsymbol{E}(\varepsilon(M_2, \Phi, N)),$$

where $\boldsymbol{E}$ denotes the expectation over the random variables Φ and N.

Boulesteix et al., The American Statistician 2015.

Sample size calculations

- Test statistic (paired t-test):

$$T = \frac{\overline{\Delta e}}{\sqrt{\frac{1}{J}\frac{1}{J-1}\sum(\Delta e(D_j) - \overline{\Delta e})^2}},$$

  where $\Delta e(D_j)$ is the difference between estimated errors of methods M_2 and M_1 in data set D_j and $\overline{\Delta e}$ is the mean over data sets.
- Power calculation for "sample size" J (number of data sets)

Boulesteix et al., The American Statistician 2015.

Inclusion criteria

Suggestion:

- Choose a set of candidate datasets, e.g. a database such as OpenML or ArrayExpress.
- Define strict inclusion criteria for datasets, e.g. $30 \leq n \leq 1000$, $p \leq 10$, etc.
- Select all datasets satisfying these criteria.
- Do not drop datasets because they yield unsatisfying results for your preferred method.
- Handle bugs (methods that do not output any result for a given dataset) adequately.

Protocols

Suggestion:

- Write a protocol describing the procedure adopted in the benchmark experiment.
- Consider including a "placebo method" in the comparison.
- Violate the protocol only in carefully justified cases.
- For example, avoid changing the primary endpoint, the competing methods, etc.

Blinding

- In clinical trials: double-blinding = the doctors do not know which treatment the patient is receiving.
- In benchmarking: double-blinding = the computational scientist first does not know which method produced which result. The decision to look for bugs in the code is not affected by the knowledge of which method produced the problematic result.

Publication bias

Publication Bias in Methodological Computational Research

Anne-Laure Boulesteix[1], Veronika Stierle[1] and Alexander Hapfelmeier[2]

[1]Department of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilian University, Munich, Germany. [2]Department of Medical Statistics and Epidemiology, Klinikum rechts der Isar Technical University of Munich, Munich, Germany.

Supplementary Issue: Statistical Systems Theory in Cancer Modeling, Diagnosis, and Therapy

ABSTRACT: The problem of publication bias has long been discussed in research fields such as medicine. There is a consensus that publication bias is a reality and that solutions should be found to reduce it. In methodological computational research, including cancer informatics, publication bias may also be at work. The publication of negative research findings is certainly also a relevant issue, but has attracted very little attention to date. The present paper aims at providing a new formal framework to describe the notion of publication bias in the context of methodological computational research, facilitate and stimulate discussions on this topic, and increase awareness in the scientific community. We report an exemplary pilot study that aims at gaining experiences with the collection and analysis of information on unpublished research efforts with respect to publication bias, and we outline the encountered problems. Based on these experiences, we try to formalize the notion of publication bias.

KEYWORDS: epistemology, publication practice, false research findings, overoptimism

Boulesteix et al., Cancer Informatics 2015.

"Neutral" comparison studies

- Publication bias, publish or perish, (subconscious) over-optimism
- Better expertise on own new method(s)

→ Biased research ("our new method performed better...")

→ Need for more neutral comparison studies!

- Benchmarking is the main goal; no new method is presented in the paper.
- Authors are—as a collective—approximately equally familiar with all considered methods.

Boulesteix et al., PLOS ONE 2013.
Boulesteix, Bioinformatics 2013.

Arguments against EBM-inspired benchmark experiments with real data

- "It all depends on datasets' characteristics"

→ personalized medicine ≈ meta-learning

- "Simulations yield better answers"

→ simulations ≈ animal research

- "EB science cannot replace experts' experience"

→ Is experience anything other than evidence informally collected over the years?

- "The substantive context has to be taken into account".

→ To which extent can decisions related to the substantive context be systematized and themselves benchmarked?

Thank you for your attention!

- Boulesteix, Wilson & Hapfelmeier, 2016. Can and should the choice of statistical methods be more "evidence-based"? In preparation.

- Boulesteix, Stierle & Hapfelmeier, 2015. Publication bias in methodological computational research. Cancer Informatics Suppl. 5:11-19.
- Boulesteix, Lauer, Hable & Eugster, 2015. A statistical framework for hypothesis testing in real data comparison studies. The American Statistician 69:201-212.
- Boulesteix, 2015. Ten simple rules for reducing overoptimistic reporting in methodological computational research. PLoS Computational Biology 11(4): e1004191.
- Boulesteix*, Lauer*, Eugster, 2013. A plea for neutral comparison studies in computational sciences. PLoS One 8(4):e61562. * both authors contributed equally to this work.
- Boulesteix, 2013. On representative and illustrative comparisons with real data in bioinformatics: comment on the letter to the editor by Smith et al. Bioinformatics 29:2664-6.