Findings from the “Methods in Research on Research” and CLEF eHealth initiatives

S88. Panel: Methods and Tools to Enhance Rigor and Reproducibility of Biomedical Research

Aurélie Névéol, PhD
LIMSI, CNRS, Université Paris Saclay
Disclosure

I and my spouse/partner have no relevant relationships with commercial interests to disclose.
Explaining the lack of reproducibility

Hypothesis: no malicious intent

- Research material is often unavailable
  - Medical corpora and other data due to confidentiality
  - Software due to commercial strategy
  - Seemingly insignificant details are left out of protocols

- Reporting bias
  - Lack of hindsight
  - Space limitation in papers
  - Novelty is valued more than reproducibility
Learning from reproducibility (or lack thereof)

• The tale of the Zigglebottom tagger

• Variability lies in…
  • Pre-processing (what is being pre-processed?)
    • Tokenization
    • Stop-word lists
    • “Data cleaning”, e.g. normalization of case, diacritics
  • Software versions, system variations, e.g. ties, random seeds
  • Parameters, including training/test split


Variability on corpus: GRACE

Counting « words »

Counting « sentences »
Improving reproducibility

• Raising community awareness
  → This panel, Pedersen CL 2008, Cohen et al. AMIA 2017, …

• Research material is often unavailable
  → Shared tasks
  → Shared datasets fostering reproducibility (e.g., Norman et al. S39)

• Reporting bias
  → Reporting guidelines
The Shared Task Model

- Primary goal is to provide a forum for direct comparison of approaches

  - Availability of shared material
  - Specific definition of a “task”
  - Corpora and annotations, split into training, development and test sets
  - Evaluation metrics and scripts
  - “Working Notes” papers describing participants’ approaches
Reproducing shared task results

• Reproducibility track at CLEF eHealth 2016
  • An automatic coding task
  • 4 analysts aim to reproduce participants runs, and baseline
  • Hypothesis: analysts use their usual work environment (vs. controlled environment)

• Overall, results can be reproduced, but…
  • Replication is not easy, even for a baseline method!
  • No single analyst was able to replicate every run
  • Documentation shortcomings reported

Levels of reproducibility

- Reproducibility of a value
  - Some experiments are not deterministic, e.g. using neural models


Image source: Tourille et al. LOUHI 2018
Levels of reproducibility

• Reproducibility of a value
  • Some experiments are not deterministic, e.g. using neural models

• Reproducibility of a finding
  • Different values obtained during iterations of an experiment may lead to the same finding, e.g.; A>B

• Reproducibility of a conclusion
  • Conclusions are inferred from findings, thus subject to interpretation

The PRIMAD model: which attributes can we “prime”? 

• Defining Types of Reproducibility
  • Platform
  • Research Objective
  • Implementation
  • Method
  • Actors
  • Data
    • Parameters
    • Input data

• What do we gain by priming one or the other?

# Types of Reproducibility and Gains

<table>
<thead>
<tr>
<th>Label</th>
<th>Data Parameters</th>
<th>Raw Data</th>
<th>Platform / Stack</th>
<th>Implementation</th>
<th>Method</th>
<th>Research Objective</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeat</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Determinism</td>
</tr>
<tr>
<td>Param. Sweep</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Robustness / Sensitivity</td>
</tr>
<tr>
<td>Generalize</td>
<td>(x)</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Applicability across different settings</td>
</tr>
<tr>
<td>Port</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Portability across platforms, flexibility</td>
</tr>
<tr>
<td>Re-code</td>
<td>-</td>
<td>-</td>
<td>(x)</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>Correctness of implementation, flexibility, adoption, efficiency</td>
</tr>
<tr>
<td>Validate</td>
<td>(x)</td>
<td>(x)</td>
<td>(x)</td>
<td>(x)</td>
<td>x</td>
<td>-</td>
<td>Correctness of hypothesis, validation via different approach</td>
</tr>
<tr>
<td>Re-use</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>Apply code in different settings, Re-purpose</td>
</tr>
<tr>
<td>Independent x (orthogonal)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>Sufficiency of information, independent verification</td>
</tr>
</tbody>
</table>
Use of Reporting Guidelines in Health Research

• Reporting Guidelines are recent tools
  • Majority have not been assessed for efficiency of reporting improvement
  • CONSORT have been shown to improve completeness of reporting
  • A systematic review reports that overall adherence to guidelines is suboptimal

• Impact of Reporting Guidelines
  • Before/After conducting a study
  • Training, Understanding, Implementing, Monitoring, Collaborating

Natural Language Processing and Reporting Guidelines

• NLP can facilitate adherence to reporting guidelines
  • Automatically assess guideline compliance
  • Match guideline item with implementation in manuscript

• Guidelines for reporting (bio)NLP research?
  • Study of 29 articles in proceedings of BioNLP 2016
  • 48% of papers provided pointers to data, 61% provided pointers to code, 21% provided pointers to both
  • Inter-rater agreement was .57 for data, .63 for code.

**Take Home Message:**
Reproducibility is hard to achieve!

- Aim at achieving reproducibility
  - Re-run, ask others to re-run
  - (Re-implement, port to different platforms)
  - Test on different data, vary parameters (and report!)

If something is not reproducible -> investigate! (you might be onto something)

- Aim for better procedures and documentation
  - Plan your research procedure: design a protocol, a data management plan
  - Document, document, document: the research process, environment, interim results, …

- Working reproducibly is good for science… and good for you!

Markowetz F. Five selfish reasons to work reproducibly. Genome Biol. 2015 Dec 8;16:274.
Thank you!

Email me at: neveol@limsi.fr

Horizon 2020 research and innovation programme: Marie Sklodowska-Curie grant agreement No 676207

CABeRneT ANR-13-JS02-0009-01

CLEF initiative