"Replicability of predictions across studies: challenges and opportunities" by Giovanni Parmigiani

discussed by M. Alfò
Dipartimento di Scienze Statistiche, Sapienza Università di Roma

Padova, 22nd September 2022
Statistical methods and models for complex data
800 years of research to understand a complex world

Reproducibility and Replicability

We will use, throughout, the definitions of *reproducibility* and *replicability* already given by prof. Altoè, just noting that

- the two concepts are closely connected,
- as one (reproducibility) may be thought as a prerequisite for the other (replicability)
- Reproducibility can be positively approached by good scientific practice
- while replicability are more inherent to the theory of statistics, machine learning and, in general, science
- as "(...) replicability is more nuanced, and in some cases a lack of replicability can aid the process of scientific discovery" (NAS *Consensus study report on Reproducibility and Replicability in Science*)

Just few words on reproducible research

Journals, scientific communities and other stakeholders may incentivate reproducibility by asking authors to share

- well documented data
- well documented/commented code, (explicit/implicit programming)
- appropriate readme files explaining how this works
- so one may expect *full bitwise reproduction* of the original results

As per the code, it should

- allow (with minimal manual changes) reproduce all figures/tables
- report transparent strategies to start/end
- contain the set seed when output relies on a RNG

Personal experience with *RR check*: after a few years of hard work by the Eds, about 60% of accepted manuscripts are *fully* reproducible, see also Mullard (2021, Nature News)

More on replicability

Once a Reproducible Research standard has been set, we may consider that replicability heavily depends on how

- we define and measure empirical evidence
- we use and interpret such a measure for the study at hand, and
- communicate it to the wider scientific community

However defined, *empirical evidence* is function of, at least

- inclusion/exclusion criteria
- observed/unobserved features
- models/methods for deriving empirical evidence
- ...

thus, it may be exceedingly *study-dependent*

Replication crisis

Probably started with

- Ioannidis (2005), Why most published research findings are false, *PLoS medicine*
- several, interesting, contributions in psychology
  - Bem (2011), Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect, *Journal of Personality and Social Psychology*,
  - Simmons et al. (2011) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant, *Psychological Science*
  - Simmons et al. (2017) False-positive citations*Perspectives on Psychological Science*

just to mention a few.

Crisis?

- Low reproducibility can exist even if there are no questionable practices involved, eg data forgery, p-hacking, etc.
- It is worth noting that *crisis* is not necessarily to be intended as a negative term [greek *krisis* (verb krisen) meaning *choice, decision*].
- We may consider this as a fundamental opportunity in making some steps forward in our approach to science.
- So, the talk by prof. Parmigiani and its focus on challenges and opportunities is very welcome.
- Here, replication is intended in the sense that empirical evidence should be, at least partially, *shared* by the (majority of) studies in a specific field.

Multistudy ensembles

- As noted by Patil and Parmigiani (2014), we need explore *generalizability* of predictions beyond the original studies
- According to Zhang et al. (2020) there is a gap in accuracy between CSV (cross-study validation) and CV (cross-validation)
- Here, replication may represent an opportunity rather than a issue, as multiple studies can be used for training
- Multistudy ensembles, as in Patil and Parmigiani (2020) or in De Vito et al. (2021) may help reduce the impact of study-specific heterogeneity on replicability
- The resulting ensemble learners may be thought to be *robust*,
- in the sense that their predictions are more likely replicated in different contexts and populations

From an empirical point of view

Some questions arise:

- Can the proposed ensemble approach be used with ordinal response, eg to predict (low, low-medium, medium, medium-high, high) risk groups?
- While ML procedures can be surely of help in predicting survival experience, or risk group labels, can we derive some more info on the role that specific genes play?
- Can the study-specific factors in De Vito et al. (2021) be, at least in some case, a source for model rethinking? May they represent interesting variation, recorded by just one (or few) studies with respect to one (or few populations)?
- When looking at ε-replicability, should we invest in choosing the value for ε or, rather, studying how replicability varies with ε?
- Can we consider some kind of *sensitivity* analysis, eg by perturbing observed data from a single study?

Thank You