Working with pre-registrations in the context of already existing data

Rafael Ahlskog

Uppsala, May 9th

Introduction

- Move in recent years (triggered by "replication crisis") toward more transparent practices around research design.

Introduction

- Move in recent years (triggered by "replication crisis") toward more transparent practices around research design.
- A methodological ideal: pre-registered studies.

Introduction

- Move in recent years (triggered by "replication crisis") toward more transparent practices around research design.
- A methodological ideal: pre-registered studies.
- Decide on crucial design choices and plan analyses before data is gathered.

Introduction

- Move in recent years (triggered by "replication crisis") toward more transparent practices around research design.
- A methodological ideal: pre-registered studies.
- Decide on crucial design choices and plan analyses before data is gathered.
- Increases confidence in results because design choices were made when still "blind" to the data.

Existing data and preregistrations

- Big problem:

Existing data and preregistrations

- Big problem: We often use data that has already been gathered by someone else – e.g. existing survey data, register data etc.

Existing data and preregistrations

- Big problem: We often use data that has already been gathered by someone else – e.g. existing survey data, register data etc.
- This poses challenges in the context of preregistering analyses:

Existing data and preregistrations

- Big problem: We often use data that has already been gathered by someone else – e.g. existing survey data, register data etc.
- This poses challenges in the context of preregistering analyses:
    1. You have to deal with the limitations that exist – can't design your ideal study.

Existing data and preregistrations

- Big problem: We often use data that has already been gathered by someone else – e.g. existing survey data, register data etc.
- This poses challenges in the context of preregistering analyses:
    1. You have to deal with the limitations that exist – can't design your ideal study.
    2. Lowers credibility of "data blindness" – in theory, you could have seen (parts of) the data already.

“Data blindness” ?

- Many ways in which your alpha can be artificially deflated:
  - Multiple hypothis testing, of course... but also

"Data blindness" ?

- Many ways in which your alpha can be artificially deflated:
  - Multiple hypothis testing, of course... but also
  - Which variable do we use?

“Data blindness” ?

- Many ways in which your alpha can be artificially deflated:
  - Multiple hypothis testing, of course... but also
  - Which variable do we use?
  - How is it defined/transformed/truncated?

"Data blindness" ?

- Many ways in which your alpha can be artificially deflated:
  - Multiple hypothis testing, of course... but also
  - Which variable do we use?
  - How is it defined/transformed/truncated?
  - Details about model specification:

“Data blindness” ?

- Many ways in which your alpha can be artificially deflated:
  - Multiple hypothis testing, of course... but also
  - Which variable do we use?
  - How is it defined/transformed/truncated?
  - Details about model specification:
    - Which variables do we include?

“Data blindness” ?

- Many ways in which your alpha can be artificially deflated:
  - Multiple hypothis testing, of course... but also
  - Which variable do we use?
  - How is it defined/transformed/truncated?
  - Details about model specification:
    - Which variables do we include?
    - What type of model?

"Data blindness" ?

- Many ways in which your alpha can be artificially deflated:
  - Multiple hypothis testing, of course... but also
  - Which variable do we use?
  - How is it defined/transformed/truncated?
  - Details about model specification:
    - Which variables do we include?
    - What type of model?
    - What type of test?

"Data blindness" ?

- Many ways in which your alpha can be artificially deflated:
  - Multiple hypothis testing, of course... but also
  - Which variable do we use?
  - How is it defined/transformed/truncated?
  - Details about model specification:
    - Which variables do we include?
    - What type of model?
    - What type of test?
    - Subgroups?

“Data blindness” ?

- Many ways in which your alpha can be artificially deflated:
  - Multiple hypothis testing, of course... but also
  - Which variable do we use?
  - How is it defined/transformed/truncated?
  - Details about model specification:
    - Which variables do we include?
    - What type of model?
    - What type of test?
    - Subgroups?
    - Etc...
- “Garden of forking paths” problem

“Data blindness” ?

- Many ways in which your alpha can be artificially deflated:
    - Multiple hypothis testing, of course... but also
    - Which variable do we use?
    - How is it defined/transformed/truncated?
    - Details about model specification:
        - Which variables do we include?
        - What type of model?
        - What type of test?
        - Subgroups?
        - Etc...
- “Garden of forking paths” problem
- Often easy to motivate just about anything post-facto!

"Data blindness" ?

- Many ways in which your alpha can be artificially deflated:
  - Multiple hypothis testing, of course... but also
  - Which variable do we use?
  - How is it defined/transformed/truncated?
  - Details about model specification:
    - Which variables do we include?
    - What type of model?
    - What type of test?
    - Subgroups?
    - Etc...
- "Garden of forking paths" problem
- Often easy to motivate just about anything post-facto!
- Makes violations of "data blindness" invariably artificially "deflate your alpha"

The problem when data exists

The problem when data exists

- You may have already seen the data in full

The problem when data exists

- You may have already seen the data in full
- You may have already seen parts of the data that are correlated with the full data

The problem when data exists

- You may have already seen the data in full
- You may have already seen parts of the data that are correlated with the full data
- Someone else may have seen the data and advised you on what to do with it / what "seems to work best"

The problem when data exists

- You may have already seen the data in full
- You may have already seen parts of the data that are correlated with the full data
- Someone else may have seen the data and advised you on what to do with it / what "seems to work best"
- Someone else may have published results with the same data

The problem when data exists

- You may have already seen the data in full
- You may have already seen parts of the data that are correlated with the full data
- Someone else may have seen the data and advised you on what to do with it / what "seems to work best"
- Someone else may have published results with the same data
- Someone else may have published results with parts of the data

The problem when data exists

- You may have already seen the data in full
- You may have already seen parts of the data that are correlated with the full data
- Someone else may have seen the data and advised you on what to do with it / what “seems to work best”
- Someone else may have published results with the same data
- Someone else may have published results with parts of the data
- Someone else may have published results with data is auto-correlated with your data

Example: Pick variable definitions from prior research!

	X_1	X_2	X_3
Y_1	0.1	0.3	0.2
Y_2	0.2	0.3	0.4
Y_3	0.1	0.2	0.5

Table: Fictional correlation table

Example: Pick variable definitions from prior research!

	X_1	X_2	X_3
Y_1	0.1	0.3	0.2
Y_2	0.2	0.3	0.4
Y_3	0.1	0.2	0.5

Table: Fictional correlation table

- If data partially overlaps, or has autocorrelation over time, merely knowing this can lead to a milder form of "p-hacking"

The central tension

- These issues give rise to a central tension that is particular to this setting:

The central tension

- These issues give rise to a central tension that is particular to this setting:
    1. You need to know about the data structure to write a detailed pre-analysis plan (since you don't "design" the data yourself) – the more you know, the fewer researcher-degrees-of-freedom you can leave on the table.

The central tension

- These issues give rise to a central tension that is particular to this setting:
    1. You need to know about the data structure to write a detailed pre-analysis plan (since you don't "design" the data yourself) – the more you know, the fewer researcher-degrees-of-freedom you can leave on the table.
    2. The more you know about the data, the less credible (under some conditions) is your data blindness.

The central tension

- These issues give rise to a central tension that is particular to this setting:
    1. You need to know about the data structure to write a detailed pre-analysis plan (since you don't "design" the data yourself) – the more you know, the fewer researcher-degrees-of-freedom you can leave on the table.
    2. The more you know about the data, the less credible (under some conditions) is your data blindness.
- Can be a difficult balance to strike.

Practical recommendations

- Don't look at the data beforehand, and simply state in the analysis plan that you haven't

Practical recommendations

- Don't look at the data beforehand, and simply state in the analysis plan that you haven't
- If you (or anyone else involved) have seen parts of the data, be open and specific about prior knowledge: what parts, how has it affected your decisions?

Practical recommendations

- Don't look at the data beforehand, and simply state in the analysis plan that you haven't
- If you (or anyone else involved) have seen parts of the data, be open and specific about prior knowledge: what parts, how has it affected your decisions?
- Build on data that you have been unable to see beforehand (because you didn't previously have access)

Practical recommendations

- Don't look at the data beforehand, and simply state in the analysis plan that you haven't
- If you (or anyone else involved) have seen parts of the data, be open and specific about prior knowledge: what parts, how has it affected your decisions?
- Build on data that you have been unable to see beforehand (because you didn't previously have access)
- Build on data that *could not* have been seen beforehand

Practical recommendations

- More generally: If you could have seen the data: simpler models and definitions are usually more credible: the more complicated your specs, the more forking paths you've already picked

Practical recommendations

- More generally: If you could have seen the data: simpler models and definitions are usually more credible: the more complicated your specs, the more forking paths you've already picked
- Plan several studies ahead!

Practical recommendations

- More generally: If you could have seen the data: simpler models and definitions are usually more credible: the more complicated your specs, the more forking paths you've already picked
- Plan several studies ahead!
  - Expensive and time consuming to order (especially) register data

Practical recommendations

- More generally: If you could have seen the data: simpler models and definitions are usually more credible: the more complicated your specs, the more forking paths you've already picked
- Plan several studies ahead!
  - Expensive and time consuming to order (especially) register data
  - Planning one study and seeing data that is also informative for future studies effectively ruins (or decreases) blinding

Practical recommendations

- More generally: If you could have seen the data: simpler models and definitions are usually more credible: the more complicated your specs, the more forking paths you've already picked
- Plan several studies ahead!
  - Expensive and time consuming to order (especially) register data
  - Planning one study and seeing data that is also informative for future studies effectively ruins (or decreases) blinding
  - Solution: plan (long) ahead – better spend an extra month or two on planning future studies, than spending additional years and 100Ks of SEK ordering multiple batches of data and compromising blinding

Conclusion

- Research always relies on trust – even with perfect pre-registration practices, we generally trust e.g. that the researchers didn't simply make up their data.

Conclusion

- Research always relies on trust – even with perfect pre-registration practices, we generally trust e.g. that the researchers didn't simply make up their data.
- There are more subtle ways than outright fraud that can distort the scientific record – awareness of these is a great first step.

Conclusion

- Research always relies on trust – even with perfect pre-registration practices, we generally trust e.g. that the researchers didn't simply make up their data.
- There are more subtle ways than outright fraud that can distort the scientific record – awareness of these is a great first step.
- Pre-registering analyses even with existing data poses unique challenges, and still relies on trust, but is arguably better than the prior status quo.

Conclusion

- Research always relies on trust – even with perfect pre-registration practices, we generally trust e.g. that the researchers didn't simply make up their data.
- There are more subtle ways than outright fraud that can distort the scientific record – awareness of these is a great first step.
- Pre-registering analyses even with existing data poses unique challenges, and still relies on trust, but is arguably better than the prior status quo.
- Transparency is key!