# Submitting to NeurIPS 2021 Pre-registration Workshop


## Abstract

Pre-registration reorders the traditional publishing life-cycle. The key idea is simple: to separate exploratory analysis from hypothesis testing. While both components are important elements of the scientific process, they must be kept distinct to ensure that the hypothesis testing remains meaningful. Pre-registration makes this separation explicit and allows the reader of a paper to clearly differentiate between the authors' predictions and exploratory analysis and "postdictions" (insights and intuition that were developed after observing the evidence).

In this tutorial, we briefly outline the motivations of the workshop and we guide the authors through the submission process.

The workshop website is `https://preregister.science`.


## 1 Motivation

As our field continues to rapidly expand in terms of number of researchers and number of new papers produced every year, it is vital to ensure scientific rigour, such that published papers can be deemed as a useful and reliable source of information for the growing audience consuming them. As observed in recent compelling position papers [5, 2, 3], the field appears to be affected by certain "troubling trends" [5] that are compromising the empirical credibility of the community at large. In machine learning, these trends materialise as several malpractices, such as:

- Evaluating a proposed method using a range of benchmarks and performance measures, but only reporting the results supporting a "positive" narrative for the submission.
- Conducting a brute-force trial-and-error approach until one combination of techniques is "state of the art" (SoTA), then structuring the writing to harmoniously present the successful set of techniques.
- Proposing a new contribution, but then evaluating it jointly with other orthogonal modifications (such as data augmentation or architecture tweaking) without appropriate ablation studies.

We suggest that these approaches are not the fault of any individual author, but rather are implicitly encouraged by the current review structure, which does not carefully distinguish between exploratory and confirmatory analysis. Since the evaluation protocols are not fixed in advance, there is considerable opportunity to search over different benchmarks, model variants and data selections until "good numbers" can be obtained. Much of this can happen subconsciously as the researchers evolve their idea and experimental protocol in parallel.

The status quo is problematic for several reasons:

- It over-inflates reviewers' expectation of the numerical improvements that should be yielded by a new method to be considered "a contribution". To compete, all authors must participate in this game to some extent (or suffer a considerable disadvantage in the review process).
- Evaluations that are conducted in this manner lose their statistical strength—the results are less likely to hold beyond the precise configuration used for reported experiments. Importantly, readers of a paper have no way to assess the number of confirmatory experiments

that were conducted before the final set are selected for publication. Note that greater computational resources thus become a significant advantage beyond the ideation and exploratory phase—holding all other variables fixed, they provide a larger set of experimental evaluations to select results from.

## 2 The pre-registration protocol

The aim of this workshop is to try and address this issue by introducing the *pre-registration* submission protocol to the machine learning community. The concept is simple: first, authors submit a proposal (the **proposal paper**, Section 3.1) before performing confirmatory experiments; then, if the paper is accepted, they conduct the proposed experiments and report their outcome in the **results paper** (Section 3.2). Importantly, at the end of the process, the overall paper is published irrespective of the results achieved.

This submission model was introduced to avoid wasteful replication of results in clinical trials [1] and it is now enjoying increasing popularity in several scientific communities [6]. It has been shown that studies following the pre-registration model tend to have better replicability [7] and present a healthy ratio of positive and negative results [4]. The workshop draws inspiration from recent papers by Gencoglu *et al.* [3] and Forde & Paganinini [2] advocating for this protocol as a potential mechanism to address the malpractices outlined in Section 1.

The hope is that this approach will nudge our community towards a different system of incentives, one that promotes scientific insights and rigorously evaluated ideas, not "state-of-the-art results at all costs" for the sake of getting a paper past peer review.

## 3 Structure

### 3.1 The *proposal paper*

The following is not a strict structure the authors are supposed to follow, but rather our suggestion. The authors are free to use a different one, provided that it follows the spirit of pre-registered studies described in Section 1 and 2.

We recommend to follow an outline comprising of four sections: *introduction*, *related work*, *methodology* and *experimental protocol*. The first three sections are unlikely to change drastically comparing to a traditional submission. However, please consider devoting particular focus to convincing the reader of the interest and novelty of the proposed approach, as conclusive empirical considerations will only be expressed after acceptance. As a guideline, please bear in mind that you will not be able to do significant edits to the proposal paper after it has been accepted – the *proposal* and *results* should "flow" once they are combined together.

A key section of a proposal paper is the **experimental protocol**. As a guiding principle, authors should aim to describe the planned experiments to the degree that a competent researcher working in the area could carry out the experiments without further communication.

Note that this is the appropriate section to describe things such as:

- Ablative studies.
- Variations that are orthogonal with the proposed idea; optimisers, types of regularisation, types/depth of architectures, ...
- Important hyper-parameters and hyper-parameter search.
- Baselines and recent methods to compare against.

A good protocol **does not mean testing all possible variations** and hyper-parameters. A good protocol is *clearly defined* and *appropriate for testing the hypotheses*.

For example, some authors may be considering 3 standard network architectures, A, B and C, to test their method. The protocol will list the 3 architectures explicitly, but the experiments may vary:

**Option 1** Exhaustive evaluation. Testing all architectures may be feasible or not, depending on the authors' resources.

**Option 2** Sequential evaluation, with the least expensive architectures being evaluated first. Not all may be evaluated in the end, making this option more feasible.

Another example is in choosing the hyper-parameters:

**Option 1** Grid search (expensive).

**Option 2** Random search, with the same computational budget given to all methods (flexible).

**Option 3** Manual search (documenting the number of trials) in Dataset 1, followed by evaluation with the same hyper-parameters in Dataset 2.

The idea is to describe the rules you are going to follow, which does not always mean being exhaustive. Note that because the protocol is fixed in advance, we expect reviewers to revise their expectations of experimental outcomes accordingly—in practice this means that fewer experiments or "weaker" improvements may be required to confirm or reject hypotheses.

We remark that **no conclusive results should be included** at this stage. However, **it is allowed to include in the proposal paper the results of preliminary experiments** that serve to validate the initial intuitions and motivate the experimental protocol. You should make clear the extent to which data has been explored. You can also speculate on what results are expected or what their significance would be for different outcomes.

### 3.2 The *results paper*

After the proposal paper has been reviewed and accepted, the authors will upload a new document containing two extra sections: experimental results and conclusion.

**Important note.** This phase will only start after the proposals have been accepted and presented at the day of the workshop. Indicatively, the deadline for the results paper will be in April or May 2022. The full paper, comprising the collation of both proposal and result, will undergo a further review to verify that the protocol has been respected and published with the PMLR journal. For reference, you can explore last year full papers at `http://proceedings.mlr.press/v148/`.

## 4 Paper format, review process and important dates

The **proposal paper** should not exceed *five* pages (*excluding* references). There will be no strict limit for the **results paper**, but as a guideline we suggest two to four pages.

Please use the NeurIPS 2021 format adapted to this workshop (which you can find at `https://preregister.science/author-kit/neurips2021_preregistration_template.zip`) for both the proposal and results papers. The results paper does not need an abstract or introduction.

## 5 Conclusion

Thank you for submitting to NeurIPS pre-registration workshop in machine learning! We hope it is going to be a fun and creative process for the authors and, overall, a useful experiment for the community.

## References


[1] Kay Dickersin and Drummond Rennie. Registering clinical trials. *Jama*, 290(4):516–523, 2003.

[2] Jessica Zosa Forde and Michela Paganini. The scientific method in the science of machine learning. *arXiv preprint arXiv:1904.10922*, 2019.

[3] Oguzhan Gencoglu, Mark van Gils, Esin Guldogan, Chamin Morikawa, Mehmet Süzen, Mathias Gruber, Jussi Leinonen, and Heikki Huttunen. Hark side of deep learning–from grad student descent to automated machine learning. *arXiv preprint arXiv:1904.07633*, 2019.

[4] Veronica Irvin and Robert M Kaplan. Likelihood of null effects of large nhlbi clinical trials has increased over time. *PLoS ONE*, 10, 08 2015.

[5] Zachary C Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341*, 2018.

[6] A. Nosek, C.R. Ebersole, A.C. DeHaven, , and D. T. Mellor. The preregistration revolution,. *PNAS*, 115(11):2600–2606, 2018.

[7] Gerard Swaen, O Teggeler, and Ludo Amelsvoort. False positive outcomes and design characteristics in occupational cancer epidemiology studies. *International journal of epidemiology*, 30:948–54, 11 2001.