HHS Public Access

Author manuscript

Psychol Aging. Author manuscript; available in PMC 2023 February 01.

Published in final edited form as:

Psychol Aging. 2022 February ; 37(1): 136–140. doi:10.1037/pag0000611.

Challenges and Opportunities in Pre-registration of Coordinated Data Analysis: A Tutorial and Template

Emily C. Willroth[1], **Eileen K. Graham**[1], **Daniel K. Mroczek**[1,2]

[1]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine

[2]Department of Psychology, Northwestern University

Abstract

The credibility revolution in social science has led to the recommendation and adoption of practices to increase the replicability of scientific findings. Many of the recommended practices, such as replication and pre-registration, present unique challenges for aging research given its reliance on long-term longitudinal data. In this tutorial, we propose pre-registered coordinated data analysis as a promising approach that involves both replication and pre-registration, but that overcomes the aforementioned challenges by using existing data. We discuss the benefits of pre-registering coordinated data analysis and provide an add-on template to be used in conjunction with existing pre-registration templates for pre-registering coordinated data analysis.

Keywords

credibility revolution; coordinated data analysis; integrative data analysis; pre-registration; replicability

The credibility revolution has led to new research practices aimed at increasing the replicability of psychological science (Angrist & Pischke, 2010; Vazire, 2018). Many of the recommended research practices, such as replication and pre-registration, present unique challenges for aging research which often requires longitudinal data collected over many years or decades. Direct replication of a study that took decades to conduct is time- and resource-intensive and is often not realistic. Similarly, pre-registration—the process of publicly documenting one's methods, analysis plans, and expected results before collecting data (van't Veer & Giner-Sorolla, 2016)—is often not feasible in its traditional form given that many longitudinal studies began decades prior to data analysis. Pre-registered coordinated data analysis provides an alternative solution that accomplishes many of the same goals as traditional pre-registration and replication, while meeting the unique needs of aging research.

In a coordinated data analysis, several independent datasets which share similar features (e.g., measurement of the same set of constructs) are used to address the same research question (Graham et al., in press; Hofer & Piccinin, 2009; Weston et al., 2020). Analyses are

Correspondence regarding this article should be directed to E.C. Willroth (emily.willroth@northwestern.edu; Hogan Biological Sciences Building 2127, Northwestern University, Evanston, IL 60202).

first conducted within each individual dataset using identical or nearly identical statistical models. Individual study analyses may sometimes be carried out by one lead analyst, but the magnitude of the analytic task as well as data sharing restrictions may instead require analyses to be carried out by multiple analysts across study sites. In the latter case, shared statistical code can be used to standardize the analysis process. Although standardized statistical code improves the reproducibility of findings and makes it easier to conduct analyses across multiple study sites, it is possible to conduct a coordinated analysis using almost any statistical software. After analyses are conducted within each individual study, the effect sizes from each study are synthesized using meta-analytic techniques. Coordinated data analysis provides multiple replications within a single project without the need to collect new data, making it feasible to replicate longitudinal findings, and in turn increasing confidence that those longitudinal findings are not due to the idiosyncrasies of a single dataset but instead are robust to heterogeneous features of several independent datasets.

Although pre-registration is not an inherent step in coordinated data analysis, pre-registration can increase the replicability of coordinated data analysis findings. First, pre-registration enhances transparency and allows researchers and readers to evaluate which aspects of the study are confirmatory and which aspects are exploratory (Nosek et al., 2019), preventing HARKing (Hypothesizing After Results are Known) (Kerr, 1998). Second, pre-registration restricts researcher degrees of freedom (e.g., running analyses with different sets of covariates and selectively reporting results), which can lead to false positive results (Wicherts et al., 2016). More broadly, this feature of pre-registration can be thought of as decision independence. Decision independence means that analytic decisions are not dependent on features of a given dataset or set of datasets (Srivastava, 2018). Notably, coordinated data analysis has been put forth as an alternative strategy for creating decision independence even in the absence of pre-registration (Srivastava, 2018) because analytic decisions are typically made in advance and applied to all of the datasets. Taken together, while both pre-registration and coordinated data analysis have the potential to increase the replicability of scientific findings, their combination is particularly beneficial.

Challenges of Pre-registering Coordinated Data Analysis

Because coordinated data analysis involves coordination across multiple datasets and often multiple analysts, many of the details that are typically included in a pre-registration are planned in advance as part of the coordinated data analysis process. For example, whereas a researcher conducting a single-study analysis can engage in data exploration relatively easily, unplanned exploratory analyses become quite complicated, and are even close to impossible, when they must be performed across 10 datasets that may be analyzed by as many different analysts. Instead, analytic decisions are typically planned in advance and then the same analytic decisions are applied across all datasets. In this regard, pre-registering a coordinated analysis is merely formalizing and publicizing planning steps that already take place. At the same time, coordinated data analysis presents the following challenges for pre-registration that researchers may not otherwise consider.

First, most existing pre-registration templates assume that the authors of the pre-registration have complete and accurate information about the data involved in the pre-registration,

but this is often not the case in coordinated data analysis. Researchers should seek data documentation for all studies in the coordinated analysis *before* pre-registering, to ensure as accurate of an understanding of data prior to devising an analytic plan as possible and to allow for detailed variable documentation in the pre-registration. However, because data documentation is often imperfect or inaccessible, certain information (e.g., response options, sample sizes, missing data) may not be known or may be incorrect at the time of pre-registration. This problem can arise in all forms of secondary data analysis but it is multiplied for coordinated data analysis by the number of datasets. To account for the possibility of incomplete or inaccurate information, researchers pre-registering a coordinated data analysis project need to incorporate contingency plans for potential unknowns in the data. For example, what will the researcher do if they expected a variable to be assessed with a Likert-type item but later find that it was assessed with a binary item? It is impossible to account for all possibilities, and some pieces of information may be known with more confidence than others. However, it is important to anticipate potential problems and surprises that are most likely to arise and pre-register contingency plans. One way to approach contingency plans is to pre-register general decision rules (e.g., we will allow for any response options that result in a continuous composite measure) as well as dataset-specific information (e.g., a list of the specific variables available in each dataset and how they will be scored). This allows the researchers to deviate from the original plan based on the pre-registered general decision rules if the data violate a researcher's expectation. Although we encourage researchers to pre-register these general decision rules, we recommend that they not be used as a replacement for dataset-specific information. Including dataset-specific information in addition to general decision rules reduces the likelihood that necessary contingencies will not be accounted for, in turn reducing the likelihood of deviation from the pre-registration.

Second, coordinated data analysis involves additional steps beyond traditional single-study designs that should be pre-registered, including dataset selection (i.e., which studies will be included in the coordinated data analysis), variable harmonization (i.e., operationalization of constructs across studies with different variables), model harmonization (i.e., model specification across studies with different variables and data structures), and results synthesis (i.e., approach to summarizing results across studies). These additional steps are not included in most existing pre-registration templates, which may lead researchers to ignore them in the pre-registration process. Because the results of a coordinated data analysis may differ depending on which specific datasets are analyzed, it is important to pre-register both the dataset inclusion criteria and the specific datasets that will be included. Once datasets are identified, the researcher must make decisions about how to harmonize the variables as well as the statistical models and these should be included in the pre-registration. Finally, there are multiple approaches to synthesizing results from a coordinated data analysis. Because these various approaches can lead to different conclusions, researchers should pre-register their plans for statistical as well as visual synthesis of results.

A Pre-registration Template for Coordinated Data Analysis

Below we provide a template for pre-registering coordinated data analysis projects (see Supplementary Online Materials for a blank copy of the template). This template is

organized according to the four steps unique to coordinated data analysis (dataset selection, variable harmonization, model harmonization, and results synthesis). Throughout, we include prompts to help researchers make contingency plans for unexpected scenarios that might arise after pre-registration (i.e., the discovery of an additional dataset or the absence of an expected variable). Because this template focuses only on the four steps that are unique to coordinated data analysis, we recommend using it as an add-on in conjunction with an existing pre-registration template, such as this template for secondary data analysis (Weston et al., 2019). For example, whereas general decision rules about how variables will be harmonized should be included in this add-on template, dataset-specific information about how variables are operationalized within each dataset can be included in an existing pre-registration template.

For each question in the template, we provide an example response based on a coordinated data analysis that our research team is currently conducting addressing the research question "Does personality predict utilization of general medical practitioners, dental care, or hospitals cross-sectionally and longitudinally?". The project used in the present illustration was pre-registered using the template for secondary data analysis, before the present add-on template was created (https://osf.io/eavkx/). Below, we illustrate how this pre-registration could have been improved using the add-on template. Of note, some of the information provided in the add-on template was not included in the original pre-registration and some information that was included in the original pre-registration was not specifically asked for in the secondary data analysis template that was used. By using a template specifically designed for coordinated data analysis, we are able to easily provide information that is particularly relevant to coordinated data analysis projects.

Part 1: Dataset Inclusion and Exclusion Criteria

a. How were datasets identified and selected? For example, did the researchers conduct a systematic search of data repositories? If so, which repositories and search terms were included in the search? Did the search process include strategies for locating grey datasets (i.e., datasets that are not located in data repositories)? If the researchers did not conduct a systematic search of relevant datasets, this should be noted.

Datasets were identified by searching studies included in the Integrative Analysis of Longitudinal Studies of Aging and Dementia on Maelstrom (https://www.maelstrom-research.org/mica/network/ialsa) for Big Five personality traits and healthcare utilization variables.

b. What are the minimum inclusion and exclusion criteria for including a dataset in the coordinated data analysis?

Datasets must include at least one Big Five personality trait (using any validated personality inventory) and at least one healthcare utilization variable assessed at the same timepoint (i.e., general medical practitioner, dental care, or hospital utilization). In addition, datasets must include age, sex, and education. Note that we will also use additional timepoints of healthcare utilization if available, but this is not a necessary criterion for inclusion. If available, we will also use measures of factors that enable healthcare utilization (i.e.,

income and insurance) and measures of factors that create need for healthcare utilization (i.e., chronic health conditions), however these are not necessary criteria for inclusion.

c. Which specific datasets meet the criteria outlined in 1b and will be included in the coordinated data analysis?

We identified twelve datasets that meet inclusion criteria: Berlin Aging Study, Berlin Aging Study II, Canberra Longitudinal Study, Health and Retirement Study, Longitudinal Aging Study Amsterdam, Long Beach Longitudinal Study, Midlife in the United States Study, SAPA Project, Swedish Adoption Twin Study of Aging, German Socioeconomic Panel Study, Veteran Affairs Normative Aging Study, Wisconsin Longitudinal Study.

d. If additional datasets, waves, or cohorts that are not named in 1c are identified that meet the criteria outlined in 1b, will they be added to the project? If yes, what is the latest stage at which additional data will be added?

Additional datasets that meet inclusion criteria will be added to the project only if they are identified prior to the meta-analysis stage. If new datasets are added, a timestamped amendment will be added to the project.

e. If datasets, waves, or cohorts identified in 1b are later found not to meet the criteria outlined in 1a, will they be dropped from the project? If yes, what is the latest stage at which data will be dropped?

If a dataset identified in 1c is found not to meet the inclusion criteria outlined in 1b, it will be dropped from the project at any stage.

Part 2: Variable Harmonization

a. For each variable in the study, please outline the degree of flexibility that you will allow in its operational definition. E.g., How much variation in the scale(s) used and/or the response options given will you allow?

Big Five personality traits can be assessed using any validated personality inventory. Response options will be allowed to vary as long as the resulting personality composite is approximately continuous (e.g., sum scores of binary items or mean scores of Likert-type items). Healthcare utilization can be assessed using any item or set of items that asks about use of general medical practitioners, dental care, or hospitals in the past X months. X will be allowed to vary across studies but must be specified. Response options will be allowed to vary but will be recoded into a binary variable indicating that the participant either utilized or did not utilize that type of healthcare. In addition to the key study variables, age, sex, education, income, health/dental insurance, and chronic conditions will be included as covariates and can be assessed using any method and any response scale.

b. For each variable in the study, please outline your harmonization plan including any data transformations. If the exact variable types and response options are not yet known or at not known with confidence, please provide contingency plans for each possible variable type and set of response options.

For personality, we will compute a composite score for each Big Five trait using the traditional scoring system for the personality inventory used. Then, we will z-score the resulting composite within each study. For healthcare utilization, we will transform all response formats into a single binary variable (utilized or did not utilize) for general medical practitioner utilization, dental care utilization, and hospital utilization, respectively. Age will be transformed into chronological years. Sex will be transformed such that 0 = female and 1 = male. Education will be transformed such that higher values indicate greater educational attainment and then z-scored within each study. Income will be z-scored within each study. For insurance, we will transform all response formats into a single binary variable (insured or not insured) for medical and dental insurance, respectively. For chronic conditions, we will compute a single count variable indicating the number of chronic health conditions from the following conditions: heart conditions, lung conditions, stroke, diabetes, cancer, and hypertension.

Part 3: Model Harmonization (i.e., Individual Study Analyses)

a. What is the optimal statistical model you will use to evaluate each hypothesis within the individual datasets?

A series of up to 15 binary logistic regressions will be used to predict three types of healthcare utilization (general medical practitioner, dental, and hospital) from each of the Big Five personality traits (a) cross-sectionally, and (b) longitudinally. Each Big Five personality trait will be assessed at baseline entered as a predictor in a separate set of regressions. In cross-sectional analyses, each of the three healthcare utilization variables will be assessed at baseline and entered as the dependent variable in a separate set of regressions. For longitudinal analyses, each of the three healthcare utilization variables will be assessed later in time and will be entered as the dependent variable in a separate set of regressions. Baseline age, sex, and education will be included as covariates in all models. Finally, in sensitivity analyses, enabling factors (i.e., income and insurance) will be included as covariates in on set of models and need factors (i.e., chronic conditions) will be included as a covariate in a separate set of models. Only one enabling factor (i.e., income *or* insurance) is required for a study to be included in the sensitivity analyses adjusting for enabling factors; however, both variables will be included as covariates when available.

b. What is the minimum viable model you will use to evaluate each hypothesis within the individual datasets?

The minimum viable model will be a cross-sectional binary logistic regression predicting one type of healthcare utilization from one Big Five trait, adjusting for age, sex, and education.

c. (optional) If the models described in 3a and 3b do not reflect the full range of models that will be evaluated, outline an organizational chart of possible models and corresponding data requirements.

Part 4: Results Synthesis and Reporting (i.e., Meta-Analyses)

a. How will parameter estimates be summarized to evaluate each hypothesis (e.g., individual study estimates only, mean weighted effect sizes, random effects or fixed effects meta-analysis)? If results will be summarized, will individual study results also be presented? How will between-study heterogeneity be evaluated, if at all?

We will use random-effects meta-analysis to calculate the overall weighted mean effect size, standard error, and 95% CIs across studies. To examine between-study heterogeneity of effects, we will calculate Cochrane's Q and I2. We will recalculate weighted effect sizes for each primary hypothesis test using a leave-one-out approach, in which the weighted effect size is recalculated 12 times excluding one study each time. This procedure will test whether the interpretation of any hypothesis test is driven by any one particularly large sample or particularly large effect.

b. Will parameter estimates from non-identical models be combined? If yes, how? If no, how will hypotheses be evaluated?

We will test four types of models: (1) cross-sectional models with demographic covariates only; (2) longitudinal models with demographic covariates only; (3) cross-sectional models with demographic covariates and enabling or need covariates; (4) longitudinal models with demographic covariates and enabling or need covariates. We will synthesize results within each type of model. This means that specific types of models might have a different number of studies contributing to the meta-analytic estimate. To evaluate whether differences across models are due to the different studies included in the meta-analysis, we will also compute meta-analytic estimates for each model using only the subset of studies that meet inclusion criteria for all models. For the sensitivity analyses adjusting for enabling factors, the key effect of personality on healthcare utilization will be summarized across studies with just income, studies with just insurance, and studies with income and insurance.

c. (optional) How will individual study results and synthesized results be plotted and/or visualized?

We will use a forest plot to display the effect sizes across studies.

Part 5: Supporting Documents (optional)

Supporting documents for this illustrated example can be found on the OSF page for the project (https://osf.io/g8vqm/).

a. Instructions for external study analysts: If more than one analyst is involved in the coordinated data analysis, general instructions should be provided to analysts regarding how to use other supporting documents. These instructions can be attached to this pre-registration.

b. Dataset construction document: Specific information about how to prepare individual datasets can be attached to this pre-registration. E.g., How should data be structured (wide versus long)? How should variables be coded (e.g., what items should be included in composites) or transformed (e.g., dummy coding, reverse coding, standardizations)?

c. Statistical code: To ensure consistency across datasets, a single set of statistical code should be applied to each dataset. Contingencies based on the availability of specific variables and/or data types can be built into the statistical code and should be automated whenever possible. This statistical code, and additional statistical code for synthesized results across datasets, can be attached to this pre-registration.

Summary and Conclusion

Pre-registration and replication are valuable tools for increasing the replicability of psychological science, but traditional approaches to replication and pre-registration present unique challenges for aging research and other fields with time- and resource-intensive data collection. Pre-registered coordinated data analysis is a solution which includes replication and pre-registration while addressing the unique nature of long-term longitudinal research. The template presented in this paper may be used in conjunction with existing pre-registration templates to pre-register coordinated data analysis projects. Pre-registered coordinated analysis is a promising approach to improving the replicability of aging research that requires relatively modest resources, and that is accessible to researchers with or without prior experience with open science practices.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Preparation of this manuscript was supported by three National Institute on Aging grants awarded to D.K. Mroczek (R01-AG018436, R01-AG067622, R01-AG064006). The project described in the present illustrated example was originally pre-registered on OSF (https://osf.io/eavkx/). We also provide links to optional supporting materials posted on the OSF project page (https://osf.io/g8vqm/).

References

Angrist JD, & Pischke JS (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. Journal of Economic Perspectives, 24, 3–30.

Graham EK, Willroth EC, Weston SJ, Muniz-Terrera G, Clouston SAP, Hofer SM, Mroczek DK, & Piccinin AM (in press). Coordinated integrative data analysis: Knowledge accumulation in lifespan developmental psychology. Psychology and Aging.

Hecht M, & Voelkle MC (2019). Continuous-time modeling in prevention research: An illustration. International Journal of Behavioral Development, 0165025419885026.

Hofer SM, & Piccinin AM (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. Psychological Methods, 14, 150. [PubMed: 19485626]

Kerr NL (1998). HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review, 2, 196–217. [PubMed: 15647155]

Nosek BA, Beck ED, Campbell L, Flake JK, Hardwicke TE, Mellor DT, … & Vazire S (2019). Preregistration is hard, and worthwhile. Trends in Cognitive Sciences, 23, 815–818. [PubMed: 31421987]

Srivastava S (2018). Sound inference in complicated research: A multi-strategy approach. 10.31234/osf.io/bwr48

van't Veer AE, & Giner-Sorolla R (2016). Preregistration in social psychology—A discussion and suggested template. Journal of Experimental Social Psychology, 67, 2–12.

Vazire S (2018). Implications of the credibility revolution for productivity, creativity, and progress. Perspectives on Psychological Science, 13, 411–417. [PubMed: 29961410]

Voelkle MC, Oud JH, Davidov E, & Schmidt P (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. Psychological methods, 17(2), 176. [PubMed: 22486576]

Weston SJ, Graham EK, & Piccinin AM (2020). Coordinated data analysis: A new method for the study of personality and health. In Personality and Healthy Aging in Adulthood (pp. 75–92). Springer, Cham.

Weston SJ, Ritchie SJ, Rohrer JM, & Przybylski AK (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. Advances in Methods and Practices in Psychological Science, 2, 214–227. [PubMed: 32190814]

Wicherts JM, Veldkamp CL, Augusteijn HE, Bakker M, Van Aert R, & Van Assen MA (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. Frontiers in Psychology, 7, 1832. [PubMed: 27933012]

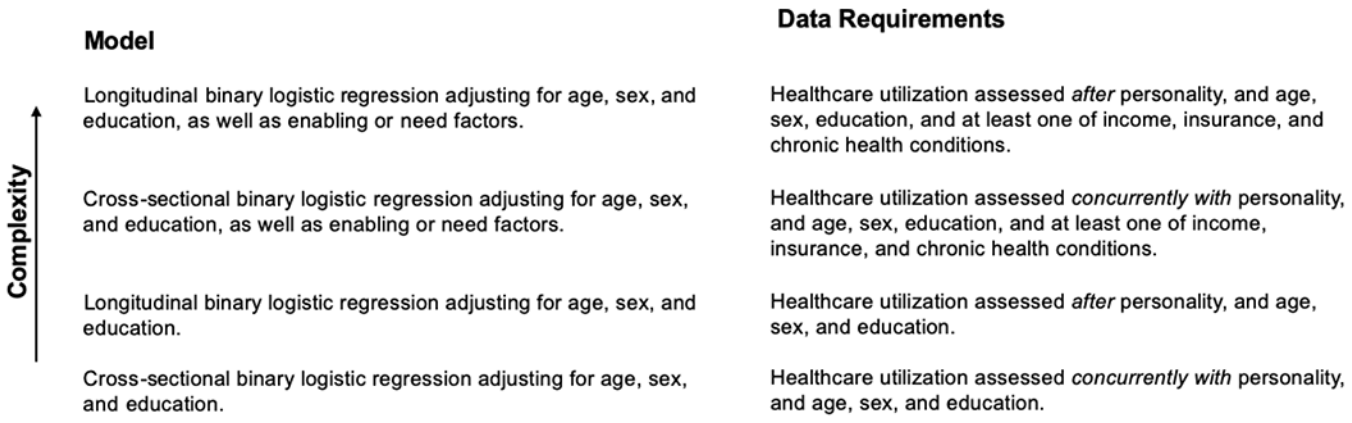

Figure 1.
Example Organizational Chart of Possible Models and Corresponding Data Requirements