Scoping Review Protocol: Statistical Models for Longitudinal Data in Health and Biomedical Research: Current State, Challenges, and Opportunities

Ariel I. Mundo Ortiz

2022-09-15

Table of contents

1 Notes

As of Sept 7, 2022 this document follows the structure recommended by PRISMA-P https://prisma-statement.org/documents/PRISMA-P-checklist.pdf

Scoping review is exploratory, can be a little broad but is best to start with one to make sure that the method works, and that its not too biased because of dispairing standards within subfields. Oncology, neurodevelopment, mental health: psichology, psychyatry

- Oncology can be the sandbox.

2 Registration

This section will be populated with the registration number and registry name once the protocol is submitted for peer review.

3 Author Contributions

- AM: Writing, query design, data extraction and analysis . . .

Other authors to add later

4 Amendements

Protocol amendments resulting from peer review will be indicated in this section indicating the date of each amendment.

5 Support

This section will indicate the sources of financial or other support for the review

5.1 Sources

6 Introduction

6.1 Rationale

Longitudinal studies are frequently used in the health sciences (biomedical research, epidemiology, public health, among others) to examine the temporal effect of a treatment or intervention **add about those studies where there is not an intervention, but follow up/evolution**[1,2]. However, the statistical analysis of longitudinal data requires to take into consideration factors such as data missingness, correlation, and non-linear trends[3,4], which represent an "analytic cost" associated with the complexity of longitudinal data[2].

One of the problems derived from the “analytic cost” of longitudinal data pertains the misspecification of the statistical models used to analyze such data (i.e., the use of models that are not coherent with the data), a problem that has been shown to occur frequently in the health sciences[5]. This problem with model misspecification can be linked to a historical preference by researchers to use the repeated measures analysis of variance (rm-ANOVA) as the default method to analyze longitudinal data, despite the fact that the multiple assumptions required by this model are frequently not satisfied by the data collected in longitudinal studies[4].

On the other hand, multiple modern statistical models were developed during the last 30 years to address the limitations of rm-ANOVA. Linear mixed models, generalized additive mixed models, and generalized estimating equations are among these modern statistical models developed for longitudinal data[6–10]. However, the use of such modern statistical methods has been the exception rather than the norm in the health sciences[11], even on this day and age where these modern methods have been brought to a wider audience with the development of computational tools such as Python or R.

Unfortunately, the misuse and lack of reproducibility of statistical analyses continue to be major problems in the health sciences[12–15]. In the case of longitudinal data, where modern methods exist beyond rm-ANOVA that can help researchers obtain better inference from their data, there is a need to understand what are the trends in the adoption of these statistical methods in the health sciences to measure the adoption of reproducibility practices by the field at large, while also identifying the reasons that may cause researchers use avoid the use of modern statistical methods for longitudinal data.

7 Objectives

This study aims to:

- Identify the different statistical models for longitudinal data that are used in the health sciences in order to measure the current extent in the adoption of modern statistical methods by the field (Aim 1a)
- Summarize the computational tools used by researchers in the health sciences to statistically analyze longitudinal data to understand the current status of the field with regards to reproducibility. (Aim 1b)
- List statistical methods for longitudinal data developed within the last decade in order to showcase

newer methods that may be applicable for longitudinal data in a biomedical/health context. (Aim 2) **maybe a different database different from Web of Science? Database for Stats or Math?**

8 Review Question

- What are the statistical methods used in biomedical/health sciences research?
- Has the use of modern statistical methods increased in the field during the last 20 years?
- What computational tools are most commonly used by researchers to analyze longitudinal data, and how in turn this affects reproducibility?
- What are most recent statistical methods developed for longitudinal data, and how can they be applied in the health sciences?

9 Methods

9.1 Types of Studies

For all the study aims, studies included in the analysis correspond to peer-reviewed publications in English.

9.2 Eligibility Criteria

9.2.1 For the Application of Modern Statistical Models on Longitudinal Biomedical/Health Data (Aims 1a and 1b)

9.2.1.1 Inclusion Criteria

Articles that are written in English, belong to the biomedical/health sciences fields, describe the collection and analysis of continous or discrete longitudinal data, indicate the statistical model used to analyze the data, and report the results of their statistical analyses.

9.2.1.2 Exclusion Criteria

Cross-sectional studies, tutorials that present the application of existing statistical methods to biomedical/health data, reviews, meta-analyses, or systematic reviews on existing statistical methods for longitudinal data, studies that use only descriptive statistics to summarize/analyze the data, studies that collect and analyze categorical data. **You don't want to exclude things right away, much rather get them and then decide.**

9.2.2 For Methods on Longitudinal Data (Aim 2)

9.2.2.1 Inclusion Criteria

- Articles that:

Are written in English, present new methodologies or significant improvements to existing methods for longitudinal data.

9.2.2.2 Exclusion Criteria

Systematic reviews, meta-analyses, or reviews of statistical methods for longitudinal data, tutorials that present the application of existing statistical methods to biomedical/health longitudinal data.

9.3 Information Sources

Studies will be retrieved from PubMed and Web of Science.

9.4 Search Strategy

PubMed and Web of Science databases will be used. Below the full search strategy for PubMed is presented for all the aims of the scoping review.

9.4.1 For the Application of Modern Models on Longitudinal Biomedical/Health Data

9.4.1.1 PubMed

9.4.1.1.1 Query:

(biomedical OR health) AND ((repeated measures) OR (longitudinal study) OR (longitudinal data)) AND ((statistical analyses) OR (statistical analysis)) NOT (Review[Publication Type] OR Meta analy*[Publication Type]) NOT ("Statistics as Topic/methods"[Majr] OR "Statistics as Topic/statistics and numerical data"[Majr] OR "Models, Statistical"[Mesh] OR "Research Design"[Mesh])

Hits: 10,972

9.4.2 For Methods on Longitudinal Data

9.4.2.1 PubMed

9.4.2.1.1 Query 1:

("Models, Statistical" [Mesh] OR "Biostatistics/methods"[Mesh]) AND ("Longitudinal Studies"[Mesh]) NOT (Review[Publication Type] OR Meta Analys*[Publication Type] OR "editorial"[Publication Type]) NOT ("survival"[Title/abstract]) NOT ("tutorial"[title/abstract] OR "orientation"[title/abstract]) NOT (Humans[Mesh] OR Adolescent [Mesh] OR Animals[Mesh])

Hits: 142

9.5 Data Collection and Analysis

9.5.1 Selection Process and Data Management

Two reviewers will independently analyze the database search results and pre-screen articles based on title and abstract content following the aforementioned inclusion/exclusion criteria. Manuscripts from the database(s) search will be stored in the Covidence platform, where duplicated entries will be removed. For articles where pre-screening inclusion (or exclusion) is unclear based on title and abstract analysis, full-text review will be used to make a decision following review by a third independent reviewer. Manuscripts included after title and abstract pre-screening will be further screening by two reviewers that will independently examine the full text of each article.

9.5.2 Data Collection Process

Pilot forms (electronic spreadsheets) will be tested using a representative sample of the studies to be reviewed (~100 studies). Information in the forms will be independently included by each reviewer. The forms will be updated (if needed), after the pilot test by consensus between the reviewers.

Information obtained from each study (statistical method used, software, etc.) will be tabulated independently by the reviewers in an electronic spreadsheet.

9.6 Data Items

Aims 1a and 1b:

Statistical method used, sub-area of application (oncology, psychology, public health, etc), computational tool used, congruence between statistical method used and the data, year of publication

Aim 2:

Statistical method reported, assumptions of the model, computational tools available for its implementation, year of publication

9.7 Risk of Bias in Individual Studies

N/A

9.8 Data Synthesis

The data from the results of each included study will be extracted into electronic spreadsheets. Summary measures for Aims 1a and 1b include plots (pie, bar, etc.) to show the relative use of each statistical method reported, computational tool, and congruence between statistical method and the data. Each plot will be segmented by year to show trends over time. Table 1 presents the headers of the pilot electronic spreadsheet.

The pilot electronic spreadsheet can be found in the following link: Pilot Spreadsheet

For Aim 2, a table will be created where statistical method, year of publication, assumptions of the model, and applicability to health data is reported.

9.9 Meta-Biases

N/A

10 References

1. Edwards LJ. Modern statistical techniques for the analysis of longitudinal data in biomedical research. *Pediatric Pulmonology*. 2000;30(4):330-344. doi:https://doi.org/10.1002/1099-0496(200010)30:4%3C330::AID-PPUL10%3E3.0.CO;2-D

2. Zeger SL, Liang K-Y. An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*. 1992;11(14-15):1825-1839. doi:https://doi.org/10.1002/sim.4780111406

3. Caruana EJ, Roman M, Hernández-Sánchez J, Solli P. Longitudinal studies. *Journal of Thoracic Disease*. 2015;7(11):E537-40.

4. Mundo AI, Tipton JR, Muldoon TJ. Generalized additive models to analyze nonlinear trends in biomedical longitudinal data using r: Beyond repeated measures ANOVA and linear mixed models. *Statistics in Medicine*. Published online July 2022.

5. Thiese MS, Arnold ZC, Walker SD. The misuse and abuse of statistics in biomedical research. *Biochem Med (Zagreb)*. 2015;25(1):5-11.

6. Linear mixed-effects models: Basic concepts and examples. In: *Mixed-Effects Models in s and s-PLUS*. Springer New York; 2000:3-56. doi:10.1007/0-387-22747-4_1

7. Jiang J, Nguyen T. *Linear and Generalized Linear Mixed Models and Their Applications*. 2nd ed. Springer; 2021.

8. Hastie TJ. *Statistical Models in S*. (Chambers JM, Hastie TJ, eds.). Routledge; 2017.

9. Rosa GJM, Gianola D, Padovani CR. Bayesian longitudinal data analysis with mixed models and thick-tailed distributions using MCMC. *Journal of Applied Statistics*. 2004;31(7):855-873.

10. Ballinger GA. Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods*. 2004;7(2):127-150.

11. Gueorguieva R, Krystal JH. Move Over ANOVA: Progress in Analyzing Repeated-Measures Data andIts Reflection in Papers Published in the Archives of General Psychiatry. *Archives of General Psychiatry*. 2004;61(3):310-317. doi:10.1001/archpsyc.61.3.310

12. Jarvis MF, Williams M. Irreproducibility in preclinical biomedical research: Perceptions, uncertainties, and knowledge gaps. *Trends in Pharmacological Sciences*. 2016;37(4):290-302. doi:https://doi.org/10.1016/j.tips.2015.12.001

13. Turkiewicz A, Luta G, Hughes HV, Ranstam J. Statistical mistakes and how to avoid them – lessons learned from the reproducibility crisis. *Osteoarthritis and Cartilage*. 2018;26(11):1409-1411. doi:10.1016/j.joca.2018.07.017

14. Gosselin R-D. Statistical analysis must improve to address the reproducibility crisis: The ACcess to transparent statistics (ACTS) call to action. *Bioessays*. 2020;42(1):e1900189.

15. Lang TA, Altman DG. Basic statistical reporting for articles published in biomedical journals: The "statistical analyses and methods in the published literature" or the SAMPL guidelines. *Int J Nurs Stud*. 2015;52(1):5-9.

Table 1: Pilot spreadsheet for data extraction

DOI	Title	Subfield	Journal	Question	Country	Source of Result (Data)	Year	Statistical Method	Software	Model assumptions checked?	Data/Model Congruency?	Code available?	Notes