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

Ariel I. Mundo Ortiz

2022-09-09

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

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) as they allow to examine how the temporal effect of a treatment or an intervention, in contrast to a cross-sectional study, which only allows to examine the effect of the intervention at a single time point. When compared their cross-sectional counterparts, longitudinal studies allow for increased statistical power and more cost efficient strategies[1,2]. However, the statistical analysis of longitudinal data requires to take into consideration factors such as data missingness, correlation, and non-linear trends,

which do not occur on cross-sectional data[3,4]. In other words, there is an "analytic cost" associated with the increased complexity of longitudinal data[2].

This additional layer of complexity has led to a problem of model misspecification in the statistical analysis of the data (i.e., the use of a statistical model that is not coherent with the data), which has been reported to occur in many fields, including the health sciences[5]. For example, in a landmark study Liu et al. showed that in a subset of papers in the biomedical sciences, the most popular model used to analyze longitudinal data was the analysis of variance (ANOVA, an approach that fails to take into account the correlation between measures over time), and that only 18% of the studies analyzed used models intended for longitudinal analysis while checking that the assumptions of the model were satisfied by the data[6].

Historically, the repeated measures ANOVA (rm-ANOVA, a statistical model for longitudinal data) has been the preferred method in the health sciences 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, the last 30 years have seen incredible progress in the field of Statistics with the development of statistical models for longitudinal data that relax the assumptions of rm-ANOVA. Linear mixed models, generalized additive mixed models, and generalized estimating equations are among these modern statistical models developed for longitudinal data[7–11]. From these statistical methods, linear mixed models and generalized estimating equations are the two classes of models that have been frequently applied to analyze longitudinal data in the health sciences during the last decade[12–14].

However, modern statistical methods that are suited to analyze longitudinal data have been the exception rather than the norm in the health sciences. In 2001, a study reported that only 30% of the clinical trials analyzed used linear mixed models to analyze their results, and that the preferred method of analysis continued to be rm-ANOVA[15] (in comparison, McCullagh and Nelder's seminal book on the generalized linear model (GLM) was published in 1989[16], and there was ongoing work on the extension of the GLM framework to the mixed model case by 1993[17]). Apart from the aforementioned study, there are not recent papers that examine the use of modern statistical methods for longitudinal data in the health sciences. Such information is critical to understand if the use of these methods has increased or decreased in the field over the last 20 years, and the reasons behind such changes.

Additionally, the reproducibility crisis is an ongoing issue in the health sciences[18,19], a major component of it being the misuse and lack of reproducibility of statistical analyses[20,21]. Despite the fact that the landscape of statistical software has vastly increased in the last decade with many statistical computational tools now available to researchers, reproducibility standards vary between each computational tool[22]. Furthermore,

there is still high variability in the amount of statistical reporting across journals[23]. Understanding what statistical computational tools are used nowadays by researchers in the health sciences can provide an assessment of the advances in the field towards research reproducibility, while identifying limitations that might still be in place.

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)

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
  - 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

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

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

9.4.1.2 Web of Science

9.4.1.2.1 Query:

(ALL=(biomedical) OR ALL=(health)) AND (ALL=(repeated measures) OR ALL=(longitudinal study) OR ALL=(longitudinal data)) AND (ALL=(statistical analyses) OR ALL=(statistical analysis)) AND (DT=(Article)) NOT (WC=(Statistics Probability) OR WC=(Mathematics)) NOT (SU=(Agriculture) AND SU=(Business Economics) AND SU=(Veterinary Sciences) AND SU=(Education Educational Research) AND SU=(Business Economics) AND SU=(Social Sciences Other Topics) AND SU=(Food Science Technology) AND SU=(Anthropology) AND SU=(Linguistics) AND SU=(Sociology) AND SU=(Criminology Penology) AND SU=(Zoology) AND SU=(Meteorology Atmospheric Sciences) AND SU=(Mathematical Methods in Social Sciences) AND SU=(Geology) AND SU=(Construction Building Technology) AND SU=(Geology) AND SU=(Religion) AND SU=(Marine Freshwater Biology) AND SU=(Operations Research Management Science) AND SU=(Fisheries) AND SU=(Metallurgy Metallurgical Engineering))

Hits: 12,458

9.4.2 For Methods on Longitudinal Data

9.4.2.0.1 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

9.4.2.1 Web of Science

(ALL=(longitudinal studies) OR ALL=(repeated measures)) NOT (TI=(survival)) AND (WC=(Statistics Probability) OR (WC=Mathematics) OR (WC=Mathematics Applied))

Hits: 8,135

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.

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. Liu C, Cripe TP, Kim M-O. Statistical issues in longitudinal data analysis for treatment efficacy studies in the biomedical sciences. Molecular Therapy. 2010;18(9):1724-1730. doi:https://doi.org/10.1038/mt.2010.127

7. 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

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

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

10. 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.

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

12. Wang M. Generalized estimating equations in longitudinal data analysis: A review and recent developments. Advances in Statistics. 2014;2014:1-11.

13. Tian Q, Qin L, Zhu W, Xiong S, Wu B. Analysis of factors contributing to postoperative body weight change in patients with gastric cancer: Based on generalized estimation equation. PeerJ. 2020;8(e9390):e9390.

14. Şevik M, Doğan M. Epidemiological and molecular studies on lumpy skin disease outbreaks in turkey during 2014-2015. Transboundary and Emerging Diseases. 2017;64(4):1268-1279.

15. 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

16. McCullagh P, Nelder JA. Generalized Linear Models. Routledge; 2019.

17. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 1993;88(421):9-25. doi:10.1080/01621459.1993.10594284

18. 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

19. 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

20. 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.

21. 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.

22. Gentleman R, Lang DT. Statistical analyses and reproducible research. Journal of Computational and Graphical Statistics. 2007;16(1):1-23. Accessed August 16, 2022. http://www.jstor.org/stable/27594227

23. Indrayan A. Reporting of basic statistical methods in biomedical journals: Improved SAMPL guidelines. Indian Pediatrics. 2020;57(1):43-48. doi:10.1007/s13312-020-1702-4