Scoping Review Protocol: Statistical Models for Longitudinal Data

Ariel I. Mundo Ortiz

2022-08-23

Table of contents

1 Background

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 models, Bayesian 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 (software, packages) 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.

In this study, we surveyed the statistical methods used in papers dealing with longitudinal data in the health sciences in order to: 1) identify statistical methods used in order to assess the trends in adoption of modern statistical methods, 2) determine what are the computational tools used by researchers to perform statistical analyses, and 3) use the previous points to provide context to the current status of the advances in research reproducibility in the field.

2 Objective

This study aims to summarize the different statistical models for longitudinal data that are used in the health sciences to identify the current extent in the adoption of modern statistical methods, determine what are the computational tools used in each case and how this in turn affects the reproducibility, and provide an updated list on methods recently developed for longitudinal data in order to determine if they can be broadly applied to longitudinal data in the health sciences.

3 Review Question

Summarize the statistical methods used to analyze longitudinal data in the health sciences to identify which methods are most commonly used, the applicability of such methods in the context of each study, and gaps that might exist that prevent the adoption of modern statistical methods that can be better suited to analyze the data. Additionally, identify if studies check for model assumptions, and how this in turn impacts the reported results.

4 Databases

- PubMed
- Web of Science

5 Search Terms

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

5.1.1 PubMed

5.1.1.1 Query 1:

(biomedical OR health) AND ((repeated measures) OR (longitudinal study) OR (ANOVA) OR (mixed effects) OR (growth curve) OR (generalized additive model) OR (generalized estimating equation)) NOT ((review) OR (meta analysis))

Hits: 393,188

Comments: query picks too many papers, and is not specific

5.1.1.2 Query 2:

(biomedical OR health) AND ((repeated measures) OR (longitudinal study)) AND ((statistical analyses) OR (statistical analysis)) NOT ((review) OR (meta analysis))

Hits: 12,617

Comments: This is the best query so far.

Papers from this query appear to be good. The query catches many papers from psychology and psychiatry, but the ones I checked did said used linear mixed models or regression in their analyses.

5.1.2 Web of Science

5.1.2.1 Query 1:

WC=(biom* OR health OR allergy OR cell biology OR cardio* OR hematology OR immunology OR life sciences biomedicine other topics OR medical informatics OR neuro* OR oncology OR pharmacology OR radiology, nuclear medicine & medical imaging OR research & experimental medicine OR substance abuse OR optics) AND AK=(longitudinal study OR repeated measures study) NOT ALL=(review OR meta analysis) NOT AK=(model* AND study design) NOT KP=(model)

Hits: 4,716

Comments: This query seems to be good.

Web of Science allows to specify more fields that result in a more targeted search. The last two parts of the query (AK and KP) removed studies method or tutorial papers from journals such as *Statistics in Medicine*.

5.2 For Methods on Longitudinal Data

5.2.1 Web of Science

5.2.1.1 Query 1:

AK=((longitudinal OR repeated measures OR longitudinal data) AND (model OR design)) NOT ALL=(review OR meta analysis) NOT ALL=(survival analysis)

Hits: 3,071

Comments: This query seems to be good.

This query returns papers that deal with methods for longitudinal analysis. Two additional options can be selected: 1) include only articles (which reduces the number of hits to 2,936 as book chapters and editorials are omitted) and 2) select from the 01/01/2000 until today (which could be reasonable as the increment of models has occurred during the last two decades. This option reduces the number to papers to 2,849).

6 Criteria for Study Selection

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

6.1.1 Inclusion Criteria

- Articles that:
  - Belong to the biomedical/health sciences fields
  - Describe the collection and analysis of longitudinal data at the preclinical or clinical level
  - Indicate the statistical model used to analyze the data
  - Report the results of their statistical analyses

6.1.2 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

6.2 For Methods on Longitudinal Data

6.2.1 Inclusion Criteria

- Articles that:

- Present new methodologies or significant improvements to existing methods for longitudinal data

6.2.2 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

7 Additional Resources

8 Comparison

- Methods most commonly used by researchers to analyze longitudinal data
- Software and packages used (R, SAS, SPSS, etc)
- Increase or decrease in the adoption of modern statistical methods for longitudinal data in the last 20 years (vs rm-ANOVA or non-parametric alternatives)
- Appropriateness of methods used in each case with regard to missing data, non-linear trends, correlation
- Articles that make clear statements about open science and that share resources (data, code, resoures sharing)

9 Data Extraction

10 Data Synthesis Strategy

11 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