Scoping Review Protocol: Statistical Models for Longitudinal Data

Ariel I. Mundo Ortiz

2022-08-12

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 to cross-sectional studies, 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 with the misspecification of the statistical models used to analyze longitudinal data (the use of a statistical model that is not coherent with the data) which has been reported in the past in the health sciences[5]. Such problem can be partially explained by the fact that researchers have a tendency to use the same statistical analyses, methods, and tests from other papers without having a clear understanding of the limitations, assumptions, and applicability of the model in each situation[5,6]. 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 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[7].

Historically, the repeated measures analysis of variance (rm-ANOVA) has been the preferred method in the health sciences to analyze longitudinal data, despite the fact that frequently, the data does not satisfy assumptions required for its use[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 overcome the limitations of rm-ANOVA. Such modern statistical models include linear mixed models, generalized additive models, Bayesian models, and generalized estimating equations among others[8–12]. However, the adoption of these modern statistical techniques has been slow within the health sciences, as showcased by Gueorguieva et al., who showed that by 2001, only 30% of clinical trials reported in the *Archives of General Psychiatry* used linear mixed models to analyze their results, and that rm-ANOVA continued to be the preferred method of analysis in most cases[13]. By comparison, McCullagh and Nelder wrote a seminal book on the generalized linear model (GLM) by 1989[14], and Breslow and Clayton reported work on the extension of the GLM framework to the mixed model case by 1993[15].

However, progress has been made during the last decade as statistical models such as generalized additive models, generalized estimating equations, and linear mixed models have been used by different groups to analyze longitudinal data in the health sciences[16–19]. Despite this, the current status in the adoption of these modern statistical methods by the field at large remains unknown. Because the use of appropriate statistical tools is a core component of research reproducibility[20,21], there is a need to better understand the current status of statistical practices for longitudinal data in the health sciences, and to identify the ongoing issues in the adoption of modern statistical models for longitudinal data.

To answer these questions, in this study we surveyed the statistical methods used in papers dealing with longitudinal data in health sciences over the last 20 years, in order to gain a better understanding of: 1) the trends in adoption of modern statistical methods, 2) identify the most frequent pitfalls in the analsys of longitudinal data, and 3) provide a rationale for situations where these methods are still not widely adopted.

2 Objective

This study aims to summarize the different statistical models for longitudinal data that are used in the health sciences, identify the extent of the adoption of modern statistical methods in the field, and determine if in each case, model assumptions are checked by researchers to ensure congruency between the data and the model.

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

6 Criteria

6.1 Inclusion Criteria

- methods paper see new methods developed
- application

6.2 Exclusion Criteria

7 Additional Resources

8 Comparison (?)

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. Ercan I, Yazici B, Yaning Y, et al. Misusage of statistics in medical research. *European Journal of General Medicine*. 2007;4(3):128-134.

7. 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

8. 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

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

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

11. 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.

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

13. 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

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

15. 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

16. Mundo AI, Muhammad A, Balza K, Nelson CE, Muldoon TJ. Longitudinal examination of perfusion and angiogenesis markers in primary colorectal tumors shows distinct signatures for metronomic and maximum-tolerated dose strategies. *Neoplasia*. 2022;32:100825. doi:10.1016/j.neo.2022.100825

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

18. 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.

19. 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.

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.