

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

4 Ariel I. Mundo Ortiz

5 2022-08-30

6 **Table of contents**

7	1 Background	2
8	2 Objective	4
9	3 Review Question	4
10	4 Databases	4
11	5 Search Terms	5
12	5.1 For the Application of Modern Models on Longitudinal Biomedical/Health Data	5
13	5.1.1 PubMed	5
14	5.1.2 Web of Science	6
15	5.2 For Methods on Longitudinal Data	7
16	5.2.1 Web of Science	7

17	6 Criteria for Study Selection	8
18	6.1 For the Application of Modern Statistical Models on Longitudinal Biomedical/Health Data	8
19	6.1.1 Inclusion Criteria	8
20	6.1.2 Exclusion Criteria	8
21	6.2 For Methods on Longitudinal Data	8
22	6.2.1 Inclusion Criteria	8
23	6.2.2 Exclusion Criteria	9
24	7 Additional Resources	9
25	8 Comparison	9
26	9 Data Extraction	9
27	10 Data Synthesis Strategy	10
28	11 References	10

29 1 Background

30 Longitudinal studies are frequently used in the health sciences (biomedical research, epidemiology, public
31 health, among others) as they allow to examine how the temporal effect of a treatment or an intervention,
32 in contrast to a cross-sectional study, which only allows to examine the effect of the intervention at a single
33 time point. When compared their cross-sectional counterparts, longitudinal studies allow for increased
34 statistical power and more cost efficient strategies^{1,2}. However, the statistical analysis of longitudinal data
35 requires to take into consideration factors such as data missingness, correlation, and non-linear trends,
36 which do not occur on cross-sectional data^{3,4}. In other words, there is an “analytic cost” associated with
37 the increased complexity of longitudinal data².

38 This additional layer of complexity has led to a problem of model misspecification in the statistical analysis
39 of the data (i.e., the use of a statistical model that is not coherent with the data), which has been reported to
40 occur in many fields, including the health sciences⁵. 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 data⁶.

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 studies⁴. 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-ANOVA¹⁵ (in comparison, McCullagh and Nelder’s seminal book on the generalized linear model (GLM) was published in 1989¹⁶, and there was ongoing work on the extension of the GLM framework to the mixed model case by 1993¹⁷). 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 tool²². Furthermore, there is still high variability in the amount of statistical reporting across journals²³. 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

72 health sciences in order to: 1) identify statistical methods used in order to assess the trends in adoption of
73 modern statistical methods, 2) determine what are the computational tools used by researchers to perform
74 statistical analyses, and 3) use the previous points to provide context to the current status of the advances
75 in research reproducibility in the field.

76 **2 Objective**

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

82 **3 Review Question**

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

88 **4 Databases**

- 89 • PubMed
- 90 • Web of Science

91 5 Search Terms

92 5.1 For the Application of Modern Models on Longitudinal Biomedical/Health 93 Data

94 5.1.1 PubMed

95 5.1.1.1 Query 1:

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

99 Hits: 393,188

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

101 5.1.1.2 Query 2:

102 (biomedical OR health) AND ((repeated measures) OR (longitudinal study) OR longitudinal data) AND
103 ((statistical analyses) OR (statistical analysis)) NOT ~~((review) OR (meta analysis))~~ when you put NOT
104 that might exclude papers with Review and meta analysis as words in the paper

105 Hits: 12,617

106 Response to comment: I followed your advice and re-wrote the query, but now I was sure to exclude meta-
107 analysis and review papers by type of publication, and papers that are classified in PubMed as devoted to
108 Statistical methodologies (not about application of methods to longitudinal data):

109 5.1.1.3 Query 3 (modified Query 2):

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

114 Hits: 10,948

115 Comments: This query is better than Query 2.

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. A few of them still deal with methodologies, but seems to be much more less than in the previous query.

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

when you put NOT that might exclude papers with Review and meta analysis as words in the paper

Remove the NOT and check the suggestions for the PubMed section

5.1.2.1.1 Query 2 (Updated 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 AK=(model* AND study design) NOT KP=(model)

Comments: Updated this query based on your comments.

Hits: 4,612

Web of Science allows to specify more fields that result in a more targeted search. I ran the search above and selected “Article” from the “Document Type” Menu. This ensures that no reviews or meta -analysis are included in the results.

140 5.2 For Methods on Longitudinal Data

141 **I re-wrote the query here based on your comments.** After reading your comments and the queries
142 I created, I realized that the filtering was not correct. I wrote a new query that I believe better represents
143 the terms we want to look at.

144 5.2.0.1 Query 1:

145 (“Models, Statistical” [Mesh] OR “Biostatistics/methods”[Mesh]) AND (“Longitudinal Studies”[Mesh])
146 NOT (Review[Publication Type] OR Meta Analys[Publication Type] OR “editorial”[Publication Type])
147 NOT (“survival”[Title/abstract]) NOT (“tutorial”[title/abstract] OR “orientation”[title/abstract]) NOT
148 (Humans[Mesh] OR Adolescent [Mesh] OR Animal[Mesh])

149 Hits: 142

150 Comments:

151 The rationale for this query is to find papers that have been labeled as dealing with models in Biostatistics
152 or Statistics, that deal with longitudinal data, but excluding reviews, editorials, meta analysis, tutorials
153 (that show how to implement an existing model, but not the development of a new model). Additionally,
154 I added the “humans”, “adolescent”, and “animal” labels to exclude, because there are **many** papers that
155 have all the previous labels but that are devoted to comparing methods, or about studies with animal or
156 clinical data (without those last filters for humans, adolescents, and animals the hits are 14,702).

157 Again, papers that describe the development of new methods for longitudinal data should be relatively few
158 when compared to papers that deal with application, and that is why to me the result of the query (142
159 hits) makes sense. I did take a look at the papers of this query and all of them seem to be about models,
160 which is what we want.

161 5.2.1 Web of Science

162 5.2.1.1 Query 1:

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

165 Hits: 3,071

166 Comments: [This query seems to be good.](#)

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

171 **6 Criteria for Study Selection**

172 **6.1 For the Application of Modern Statistical Models on Longitudinal Biomed-** 173 **ical/Health Data**

174 **6.1.1 Inclusion Criteria**

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

180 **6.1.2 Exclusion Criteria**

- 181 • Cross-sectional studies
- 182 • Tutorials that present the application of existing statistical methods to biomedical/health data
- 183 • Reviews, meta-analyses, or systematic reviews on existing statistical methods for longitudinal data
- 184 • Studies that use only descriptive statistics to summarize/analyze the data

185 **6.2 For Methods on Longitudinal Data**

186 **6.2.1 Inclusion Criteria**

- 187 • Articles that:
 - 188 – Present new methodologies or significant improvements to existing methods for longitudinal data

189 6.2.2 Exclusion Criteria

- 190 • Systematic reviews, meta-analyses, or reviews of statistical methods for longitudinal data
- 191 • Tutorials that present the application of existing statistical methods to biomedical/health longitudinal
- 192 data

193 7 Additional Resources

194 8 Comparison

- 195 • Methods most commonly used by researchers to analyze longitudinal data
- 196 • Software and packages used (R, SAS, SPSS, etc)
- 197 • Increase or decrease in the adoption of modern statistical methods for longitudinal data in the last
- 198 20 years (vs rm-ANOVA or non-parametric alternatives)
- 199 • Appropriateness of methods used in each case with regard to missing data, non-linear trends, corre-
- 200 lation
- 201 • Articles that make clear statements about open science and that share resources (data, code, resources
- 202 sharing)

203 9 Data Extraction

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

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](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](https://doi.org/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.

- 237 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.
238 2020;8(e9390):e9390.
- 239 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.
- 241 15. Gueorguieva R, Krystal JH. Move Over ANOVA: Progress in Analyzing Repeated-Measures Data
and Its Reflection in Papers Published in the Archives of General Psychiatry. Archives of General
242 Psychiatry. 2004;61(3):310-317. doi:[10.1001/archpsyc.61.3.310](https://doi.org/10.1001/archpsyc.61.3.310)
- 243 16. McCullagh P, Nelder JA. Generalized Linear Models. Routledge; 2019.
- 244
- 245 17. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the
246 American Statistical Association. 1993;88(421):9-25. doi:[10.1080/01621459.1993.10594284](https://doi.org/10.1080/01621459.1993.10594284)
- 247 18. Jarvis MF, Williams M. Irreproducibility in preclinical biomedical research: Perceptions, uncer-
tainties, and knowledge gaps. Trends in Pharmacological Sciences. 2016;37(4):290-302. doi:<https://doi.org/10.1016/j.tips.2015.12.001>
- 248
- 249 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.
250 doi:[10.1016/j.joca.2018.07.017](https://doi.org/10.1016/j.joca.2018.07.017)
- 251 20. Gosselin R-D. Statistical analysis must improve to address the reproducibility crisis: The ACcess to
252 transparent statistics (ACTS) call to action. Bioessays. 2020;42(1):e1900189.
- 253 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
254 Stud. 2015;52(1):5-9.
- 255 22. Gentleman R, Lang DT. Statistical analyses and reproducible research. Journal of Computational
256 and Graphical Statistics. 2007;16(1):1-23. Accessed August 16, 2022. <http://www.jstor.org/stable/27594227>
- 257 23. Indrayan A. Reporting of basic statistical methods in biomedical journals: Improved SAMPL guide-
258 lines. Indian Pediatrics. 2020;57(1):43-48. doi:[10.1007/s13312-020-1702-4](https://doi.org/10.1007/s13312-020-1702-4)