Data Analysis Errors: Examples of Actual Mistakes in Scientific Publications

Marina Prändl (401045975)

Fresenius University of Applied Science

Data Analysis for Decision-Making (WS 2025/26)

Prof. Dr. Stephan Huber

2025-12-17

Author Note

Correspondence concerning this article should be addressed to Marina Prändl (401045975), Email: praendl.marina@stud.hs-fresenius.de

Abstract

Data analysis begins with collecting, cleaning, and structuring raw data, then applying statistical methods to interpret it. Its main purposes are either discovering new scientific knowledge or optimizing large-scale decisions such as identifying the most profitable products.To illustrate the impact of errors, an R code example using the tidyverse and babynames packages demonstrates how duplicates in a dataset can distort results and lead to inaccurate conclusions. The error is simulated by intentionally duplicating entries and then corrected using the distinct () function, which removes duplicate rows. Beyond this example, several real-world cases highlight the consequences of mistakes in scientific publications: a programming error in Springer Nature's system switch misallocated citation credit, statistical errors remained unchecked due to data withholding, and surveillance data was misinterpreted as causal proof, leading to stricter publication rules. The COVID-19 Excel error in the UK shows how technical problems can seriously affect public health, leading to thousands of cases being missed and slowing down contact tracing. These examples emphasize the importance of strict data handling, transparency, and reliable tools to prevent distorted outcomes and maintain trust in scientific and public decision-making.

Data Analysis Errors: Examples of Actual Mistakes in Scientific Publications

Word count: 1742

1 Introduction: Why Data Analysis matters

The starting point of data analysis is data collection. Once collected, the raw data must be cleaned and structured by using specific methods. The analysis then consists of applying methods like statistics to figure out what the data shows. Data analysis usually has one of the following motivations: 1. Discovering new knowledge for scientific advancement e.g. research for prevention of cancer. 2. Repetitive large-scale decisions for optimization e.g. a company analysing the sales of their product range to find the most profitable one (see Provost and Fawcett (2013)).

2 Showcasing Example Error with R Code

Assuming Germany for example would want to know which babyname is the most frequent one in their country, an analysis would be required to find an accurate result.

2.1 Step 1: Load packages

Since this is only an example to show the impact that a small error can have, an easy to follow data package is used. Therefore tidyverse and the data set package babynames requires loading. This data set contains baby name data.

```
library(tidyverse)
library(babynames)
  )
```

2.2 Step 2: Create small data set to visualise example

To start off, a data set is required. For that a small and easily understandable dataset is created. The initial dataset is called starting_dataset. <- tells R to assign the babynames from the package to the dataset, so that the entries will be filled with the information from the babynames package. %>% the pipe operator makes the code easier to read by forwarding the object prior to it, to the object after it. filter() keeps only rows where the year is 2000 and the sex is female. select() chooses only those three columns. slice() takes only the first 10 rows of the already filtered dataset.

```
# Select 10 entries: year, name, frequency
starting_data <- babynames %>%
  filter(year == 2000, sex == "F") %>%
  select(year, name, frequency = n) %>%
  slice(1:10)
```

2.3 Step 3: Add duplicates to show potential error

As this example only simulates an error, it must firstly be implemented which is done by adding duplicates. <- bind_rows() uses the starting_data as a baseline. Furthermore %>% slice() tells R to keep the first and second row besides the rest of the starting_data. As a result there are two babynames duplicated, shown in data_with_duplicates.

```
# Add duplicates: intentionally repeat 2 names
data_with_duplicates <- bind_rows(
  starting_data,
  starting_data %>% slice(1), # first name duplicated
  starting_data %>% slice(2) # second name duplicated
  )

# Show dataset with duplicates
data_with_duplicates
```

2.4 Step 4: Removing Error

This step is about correcting the error once it already has been committed. With the function distinct() using the data_with_duplicates it creates a new data set where all the names that appear double are removed. Distinct only keeps unique rows. Clean_data shows the fixed data set without any errors.

```
# Remove duplicates
clean_data <- distinct(data_with_duplicates)

# Show dataset without duplicates
clean_data
```

This example showcases that if analysts would work with a data set that contains duplicates, without noticing or correcting it they could come to a distorted and inaccurate conclusion. # Example Cases of Actual Mistakes in Scientific Publications This section provides three examples of Actual Mistakes in Scientific Publications. The respective error, causes and consequences are highlighted. For more details about the situation background (see Joelving (2025); Attride (2024); Travis (2025)).

2.5 Case Example: Programming Error in Springer Nature system switch

The core error was a Systemic Misallocation of Citation Credit, meaning the count for a reference to one paper was incorrectly recorded in the database and assigned to a different one. This happened due to a Programming Bug During Metadata Transition. The publisher changed how papers were identified, moving from using page numbers to a unique Article Number. A software mistake during this change caused many of the incoming citations to be wrongly directed to the first article (Article Number 1) in each journal volume, rather than the correct paper. The consequence is Unfairly Boosted Success Scores. The Article Number 1 papers received thousands of citations that did not belong to them, which falsely boosted their importance. This led to a Distorted Scientific Evaluation, giving researchers with these inflated counts an unfair advantage in getting jobs, grants, and promotions. Ultimately, it also caused an Erosion of Trust because it made the crucial metrics for judging research impact unreliable (see Joelving (2025)).

Solution: Provide better Quality control and software testing before switching such impactful systems to prevent programming bugs.

2.6 Case Example: Statistical Errors Unchecked Due to Data Withholding

The error was the failure to uphold scientific transparency and reproducibility that both the authors and the journal editor refused to release the raw research data when asked for verification. This stopped the wider scientific community from being able to check or correct the published claims. The cause was a deliberate Refusal of Data Sharing. The authors of the paper were raising legal issues, and the journal did not want to get involved in a legal dispute. The journal considered the matter closed because the paper had already gone through a full review process. The consequence was damaged trust in the publication. Because the data was kept secret, the original allegations of statistical errors could neither be confirmed nor disproven, meaning the paper's conclusions remain doubtful and damaging the credibility of the research, the authors, and the journal's commitment to open science (see Attride (2024)).

Solution: Uphold scientific transparency and reproducibility by pushing raw data releases. Even if data sharing raises legal concerns, prioritize the accuracy and credibility of scientific work over the individual reputation of an author or journal.

2.7 Case Example: Systemic Misinterpretation of Surveillance Data as Causal Proof

The error involves publishing research papers that use basic, unverified drug safety reports to suggest a strong link between a medicine and a side effect, even though that data is not meant to prove a cause-and-effect relationship. This happens because the literature is being overwhelmed by lot of Low-Quality, Simple Analyses. A huge number of research papers are being published using these easy-to-access safety databases. These papers often show only simple connections that lack the detailed analysis needed to confirm they are real. The consequence is a Banning of Database-Only Papers to Preserve Quality. The journal, British Journal of Clinical Pharmacology (BJCP), implemented an immediate and strict rule to reject any paper based only on these drug surveillance databases. This extreme action was taken to stop the flood of low-quality, misleading research and protect the journal's scientific standards (see Travis (2025)).

Solution: Establish stricter publication rules to reduce low quality and misleading research.

3 COVID-19 Excel Error in UK

In September 2020, health authorities in England relied on Excel spreadsheets to manage the data for COVID-19 cases and conduct contact tracing—a crucial public health tool used to quickly inform people exposed to the virus (like in other epidemics e.g. Ebola).

The system failed because the Excel spreadsheet hit its maximum row limit. A coding error meant that when the sheet ran out of space, the data for new cases was simply cut off and lost instead of generating a clear error warning. Due to this glitch, the data for 15,841 positive COVID-19 cases (about 20% of all cases at the time) was left out of official reports. The infected people were informed of their positive results, but the missing data was not passed to the contact tracing system. This meant that an estimated 48,000 contact persons could not be informed to isolate themselves through self-quarantine. This significantly slowed down the process: normally, 96% of cases were passed to tracing within five days, and 80% of contacts were reached within 24 hours. When the Excel error happened, only 60% of contacts were reached within 24 hours.

Researchers used this random error as a "natural experiment" to measure the impact of

delayed contact tracing without having to deliberately harm people. By looking at which local areas were most severely affected by the late reporting, they could compare regions reliably. They used a statistical method called "difference-in-differences" to separate the effect of the Excel glitch from the general rise in cases across the country. They found that before the glitch, all areas had similar COVID-19 trends, but the areas with the most delays saw a much stronger increase in infections afterward.

Overall, the researchers estimated that because of the Excel error, England saw between 126,000 and 185,000 extra infections and between 1,500 and 2,000 extra deaths in the six weeks following the glitch.

The error was missed because Excel lacks clear warnings: the file still opens, and data appears normal, as no obvious error message alerts users that data rows have been cut off. The missing cases also blended in with the system's natural delay (2–5 days) in processing test results, making the sudden jump in reported cases after the correction the only clear sign of the mistake (see Fetzer and Graeber (2021)).

Solution: Implement systems that provide clear error warnings. Choose reliable tools that are limitless or ensure sufficient capacity for your activities.

4 Additional Readings

Covid Data (e.g. cases, testing, deaths etc.) from the UK is provided online by (UKHSA (2025)). The files available to download, can give an insight into real large datasets. This website (Staff (2025)) shows an example of possible support by certain initiatives that fund to improve the quality and reliability of scientific work.

5 Affidavit

I hereby affirm that this submitted paper was authored unaided and solely by me. Additionally, no other sources than those in the reference list were used. Parts of this paper, including tables and figures, that have been taken either verbatim or analogously from other works have in each case been properly cited with regard to their origin and authorship. This paper either in parts or in its entirety, be it in the same or similar form, has not been submitted to any other examination board and has not been published. I acknowledge that the university may use plagiarism detection software to check my thesis. I agree to cooperate with any investigation of suspected plagiarism and to provide any additional information or evidence requested by the university.

Checklist:

- ☒ The handout contains 3-5 pages of text.
- ☒ The submission contains the Quarto file of the handout.
- ☒ The submission contains the Quarto file of the presentation.
- ☒ The submission contains the HTML file of the handout.
- ☒ The submission contains the HTML file of the presentation.
- ☒ The submission contains the PDF file of the handout.
- ☒ The submission contains the PDF file of the presentation.
- ☒ The title page of the presentation and the handout contain personal details (name, email, matriculation number).
- ☒ The filled out Affidavit.
- ☒ The handout contains a bibliography, created using BibTeX with an APA citation style.
- ☒ Either the handout or the presentation contains R code that proofs the expertise in coding.
- ☒ The link to the presentation and the handout published on GitHub.
- ☐ In group work,each student's contribution is clearly defined, and individual performance can be assessed using specified sections, page numbers, or other objective criteria.

Marina Prändl, 12/17/2025, Cologne

6 References

Attride, D. (2024). *Editor and authors refuse to share data of paper containing alleged statistical errors*.

Fetzer, T., & Graeber, T. (2021). Measuring the scientific effectiveness of contacttracing: Evidence from a natural experiment. *Proceedings of the National Academy of Sciences*.

Joelving, F. (2025). *Bug in springer nature metadata may be causing "significant, systemic" citation inflation*.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*.

Staff, R. W. (2025).

Travis, K. (2025). *Exclusive: Journal bans drug safety database papers as they flood the literature*.

UKHSA. (2025). *COVID-19 archive data download*.