Dossier: visualizing/ understanding decision choices in data analysis via decision similarity

ANONYMOUS AUTHOR(S)

In data analysis, analysts are expected to clearly communicate the decisions they make, as these choices inform how results are interpreted and compared across studies. Such decisions – for example, selecting the degree of freedom for a smoothing spline – are often not systematically studied, since onece an analysis is published, it is done seldom revisited or replicated with alternative choices. In this work, we focus on a body of data analysis studies on the effect of particulate matter on mortality, conducted by researchers worldwide, which naturally provide alteranative analyes of the same question. We automatically extract analytic decisions from the published literature into structured data using Large Language Models (Claude and Gemini). We then proposed a pipeline to calculate paper similarity based on the semantic similarity of these extracted decisions and their reasons, and visualize the results through clustering algorithms. This approach offers an efficient way to study decision-making practices than traditional interviews. We also provide insights into the use of LLMs for text extraction tasks and the communication of analytic choices in data analysis practice.

CCS Concepts: • **Applied computing** → *Document analysis*; • **Human-centered computing** → **HCI theory, concepts and models**.

Additional Key Words and Phrases: Large language models

ACM Reference Format:
Anonymous Author(s). 2025. Dossier: visualizing/ understanding decision choices in data analysis via decision similarity. In *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI'26).* ACM, New York, NY, USA, 19 pages. https://doi.org/XXXXXXX.XXXXXXX

- Something about "analysis review" - Roger thinks it's a better to have a new word for this.
- demonstrate - analytically homogeneous - the table won't look like that

1 Introduction

Decisions are everywhere in data analysis, from the initial data collection, data pre-processing to the modelling choices. These decisions will impact the final output of the data analysis, which may lead to different conclusions and policy recommendations. When such flexibility can be misused—through practices such as p-hacking, selective reporting, or unjustified analytical adjustments—it can inflate effect sizes or produce misleading results that meet conventional thresholds for statistical significance. They have been demonstrated through many-analysts experiments, where independent teams analyzing the same dataset to answer a pre-defined research question often arrive at markedly different conclusions. These practices not only compromise the validity of individual studies but also threaten the broader credibility of statistical analysis and scientific research as a whole.

Multiple recommendations have been proposed to improve data analysis practices, such as pre-registration and multiverse analysis. Bayesian methods also offer a different paradigm to p-value driven inference for interpreting statistical evidence. Most empirical studies of data analysis practices focus on specially designed and simplified analysis

Manuscript submitted to ACM

scenarios. While informative, these setups may not adequately capture the complexity of the data analysis with significant policy implications. [In practice, studying the data analysis decisions with actual applications is challenging.] Analysts may no longer be available for interviews due to job changes, and even when they are, recalling the full set of decisions and thinking process made during the analysis is often infeasible. Moreover, only until the last decades, analysis scripts and reproducible materials were not commonly required by journals for publishing. [As a result, it remains challenging to study how analytical decisions are made.]

In this work, we focus on a specific class of air pollution modelling studies that estimate the effect size of particulate matter (PM2.5 or PM10) on mortality, typically using Poisson regression or generalized additive models (GAMs). While individual modelling choices vary, these studies often share a common structure: they adjust for meteorological covariates such as temperature and humidity, apply temporal or spatial treatments, like including lagged variables and may estimate the effect by city or region before combining results. Because these studies investigate similar scientific questions using a shared modelling framework, they form a natural many-analyst setting. This allows us to examine, in a real-world context, the range of analytical decisions made by different researchers addressing the same underlying question.

In this work, we develop a structured tabular format to record the analytical decisions made by researchers in the air pollution modelling literature. Using large language models (LLMs), we automate the extraction of these decisions from published papers. This allows us to treat decisions as data – allowing us to track them over time, compare methodology across papers, and query commonly used approaches. We further introduce a workflow to cluster studies based on decision similarity, revealing three distinct groups of papers that reflect the modelling strategies differs in the European and U.S. studies, which offers a new way to visualize the field in the air pollution mortality modelling.

The contribution of this work includes:

- A new approach to study data analysis decision choices through automatic extraction of decisions from scientific literature using LLMs,
- A dataset compiled from 62 papers to study decision-making in air pollution mortality modelling,
- A pipeline to construct similarities between papers based on decision similarities, and
- Issues we found from existing data analysis reporting

The rest of the paper is organized as follows. In Section 2, we review the background on data analysis decisions. Section 3 describes the data structure for recording decisions, the use of large language models to process research papers, and the validation of LLM outputs. In Section 4, we present the method for calculating paper similarity based on decision similarities. Section 5 reports the finding of our analysis, including the clustering of papers according to similarity scores and sensitivity analyses related to LLM providers, prompt engineering, and LLM parameters. Finally, Section 7 discusses the implications of our study.

2 Related work

2.1 Decision-making in data analysis

A data analysis is a process of making choices at each step, from the initial data collection to model specification, and post-processing. Each decision represents a branching point where analysts choose a specific path to follow, and the vast number of possible choices analysts can take forms what Gelman and Loken [18] describe as the "garden of forking paths". While researchers may hope their inferential results are robust to the specific path taken through the garden, in practice, different choices can lead to substantially different conclusions. This has been empirically demonstrated through "many analyst experiments", where independent research groups analyze the same dataset to the same answer using their chosen analytic approach. A classic example is Silberzahn et al. [41], where researchers reported an odds ratio from 0.89 to 2.93 for the effect of soccer players' skin tone on the number of red cards awarded by referees. Similar variability has been observed in structural equation modeling [39], applied microeconomics [22], neuroimaging [8], and ecology and evolutionary biology [19].

Examples above have rendered decision-making in data analysis as a subject to study in data science. To collect data on how analysts making decisions during data analysis, researchers have conducted interviews with analysts and researchers on data analysis practices [2, 24, 29], visualization of the decision process through the analytic decision graphics (ADG) [30]. Recently, Simson et al. [42] describes a participatory approach to decisions choices in fairness ML algorithms. Software tools have also developed to incorporate potential alternatives in the analysis workflow, including the `DeclareDesign` package [7] and the `multiverse` package [38]. The `DeclareDesign` package [7] introduces the MIDA framework for researchers to declare, diagnose, and redesign their analyses to produce a distribution of the statistic of interest, which has been applied in the randomized controlled trial study [6]. The `multiverse` package [38] provides a framework for researchers to systematically explore how different choices affect results and to report the range of plausible outcomes that arise from alternative analytic paths. Other systems have been developed to visualize `multiverse` analysis [31].

2.2 Visualization on scientific literature

Much of the work on IEEE visualizing scientific literature focuses on helping researchers stay aware of relevant publications, given the rapidly growing volume of scientific output and the difficulty of navigating it. Systems have been developed to support the discovery of relevant papers, where relevance is typically determined by keywords [23], citation information (e.g. citation list, co-citation) [13], or combinations with other relevant paper metadata (e.g. author, title) [5, 14, 17, 20]. More recent approaches incorporate text-based information from the abstract or sections of the paper to [obtain a better similar metric]. This includes using topic modelling [1], argumentation-based information retrieval [43], and text embedding [36]. While these metadata information and high level text-based information are valuable for discovering relevant papers, for data analysis, researchers need tools that help them *make sense* of the literature rather than simply *finding* it. Capturing the decisions and reasoning expressed during analyses within a similar theme can reveal common practices in the field and guide decisions choices in new applications. With recent advances in Large Language Models (LLMs), it has become possible to automatically extract structured information from unstructured text through prompting. This allows scientific literature to be clustered and visualized using information about the underlying decisions and reasoning made during analysis, providing a basis for studying analysts' decision choices.

3 Methods

3.1 Decisions in data analysis

Decisions occur throughout the entire data analysis process – from the selection of variables and data source, to pre-processing steps to prepare the data for modelling, to the model specification and variable inclusion. In this work, we focus specifically on modelling decisions in the air pollution mortality modelling literature. These include the choice of modelling approach, covariate inclusion and smoothing, and specifications of spatial and temporal structure. Consider the following excerpt from Ostro et al. [37]:

> Based on previous findings reported in the literature (e.g., Samet et al. 2000), the basic model included a smoothing spline for time with 7 degrees of freedom (df) per year of data. This number of degrees of freedom controls well for seasonal patterns in mortality and reduces and often eliminates autocorrelation.

This sentence encode the following components of a decision:

- **variable**: time
- **method**: smoothing spline
- **parameter**: degree of freedom (df)
- **reason**: Based on previous findings reported in the literature (e.g., Samet et al. 2000); This number of degrees of freedom controls well for seasonal patterns in mortality and reduces and often eliminates autocorrelation.
- **decision**: 7 degrees of freedom (df) per year of data

The decision above is regarding a certain parameter in the statistical method, we categorize this as a "parameter" type decisions. Other types of decisions - such as spatial modelling structure or the inclusion of temporal lags - may not include an explicit method or parameter, but still reference a variable and rationale, which we will provide further examples below.

To record these decisions, we follow the tidy data principle [45], where each variable should be in a column, each observation in a row. In our context, each row represents a decision made by the authors of a paper and an analysis often include multiple decisions. To retain the original context of the decision, we extract the original text in the paper, without paraphrase or summarization, from the paper. Below we present an example of how to structure the decisions made in a paper, using the paper by Ostro et al. [37]:

Paper	ID	Model	variable	method	parameter	type	reason	decision
ostro	1	Poisson regression	temperature	smoothing spline	degree of freedom	parameter	NA	3 degree of freedom
ostro	2	Poisson regression	temperature	smoothing spline	degree of freedom	temporal	NA	1-day lag
ostro	3	Poisson regression	relative humidity	LOESS	smoothing parameter	parameter	to minimize Akaike's Information Criterion	NA
ostro	4	Poisson regression	model	NA	NA	spatial	to account for variation among cities	separate regression models fit in each city

Most decisions in the published papers are not explicitly stated, this could due to the coherence and conciseness of the writing or authors' decision to include only necessary details. Here, we identify a few common anomalies where decisions may be combined or omit certain fields:

1. **Authors may combine multiple decisions into a single sentence** for coherence and conciseness of the writing. Consider the following excerpt from Ostro et al. [37]:
   Other covariates, such as day of the week and smoothing splines of 1-day lags of average temperature and humidity (each with 3 df), were also included in the model because they may be associated with daily mortality and are likely to vary over time in concert with air pollution levels.

This sentence contains four decisions: two for temperature (the temporal lag and the smoothing spline parameter) and two for humidity. These decisions should be structured as separate entries.

2. **The justification does not directly address the decision choice.** In the example above, the stated rationale ("and are likely to vary over time in concert with air pollution levels") supports the general inclusion of temporal lags but does not justify the specific choice of 1-day lag over alternatives, such as 2-day average of lags 0 and 1 (lag01) and single-day lag of 2 days (lag2). As such, the reason field should be recorded as NA.
3. **Some decisions may be omitted because they are data-driven**. For instance, Katsouyanni et al. [26] states:
   The inclusion of lagged weather variables and the choice of smoothing parameters for all of the weather variables were done by minimizing Akaike's information criterion.

In this case, while the method of selection (minimizing AIC) is specified, the actual degree of freedom used is not. Such data-driven decisions may be recorded with "NA" in the decision field, but the reason field should still be recorded as "by minimizing Akaike's information criterion"

4. **Information required to interpret the decision may be distributed across multiple sections**. In the previous example, "weather variables" refers to mean temperature and relative humidity, as defined earlier in the text. This requires cross-referencing across sections to identify the correct variables associated with each modeling choice.

3.2 Automatic reading of literature with LLMs

TODO: Prompt engineering: these models may paraphrase or hallucinate unless explicitly told not to since it is generative in nature based on the predicted probability of the next word from the text and the instruction

TODO: The Prompt Report: A Systematic Survey of Prompt Engineering Techniques https://arxiv.org/pdf/2406.06608

While decisions can be extracted manually from the literature, this process is labor-intensive and time-consuming. Recent advances in Large Language Models (LLMs) have demonstrated potential for automating the extraction of structured information from unstructured text [ref]. In this work, we use LLMs to automatically identify decisions made by authors during their data analysis processes.

Text recognition from PDF document relies on Optical Character Recognition (OCR) to convert scanned images into machine-readable text – capability currently offered by Antropic Claude and Google Gemini. We instruct the LLM to generate a markdown file containing a JSON block that records extracted decisions, which can then be read into statistical software for further analysis. The exact prompt feed to the LLM is provided in the Appendix. The `ellmer` package [46] in R is used to connect to the Gemini and Claude API, providing the PDF attachment and the prompt in a markdown file as inputs.

3.3 Review the LLM output

- **TODO** something about result validation of LLM output: We also observe data quality with the extraction: for example in Lee et al. [28], the variable recorded is “smoothing parameter”. Authors are unclear about the delivery Specify how much of validation and review has been done.

The shiny app is designed to provide users a visual interface to review and edit the decisions extracted by the LLM from the literature. The app allows three actions from the users: 1) *overwrite* – modify the content of a particular cell, equivalently `dplyr::mutate(xxx = ifelse(CONDITION, "yyy" , xxx))`, 2) *delete* – remove a particular cell, `dplyr::filter(!(CONDITION))`, and 3) *add* – manually enter a decision, `dplyr::bind_rows()`. Figure 1 illustrates the *overwrite* action in the Shiny application, where users interactively filter the data and preview the rows affected by their edits—in this case, changing the model entry from “generalized additive Poisson time series regression” to the less verbose “Poisson regression”. Upon confirmation, the corresponding `tidyverse` code is generated, and users can download the edited table and incorporate the code into their R script.

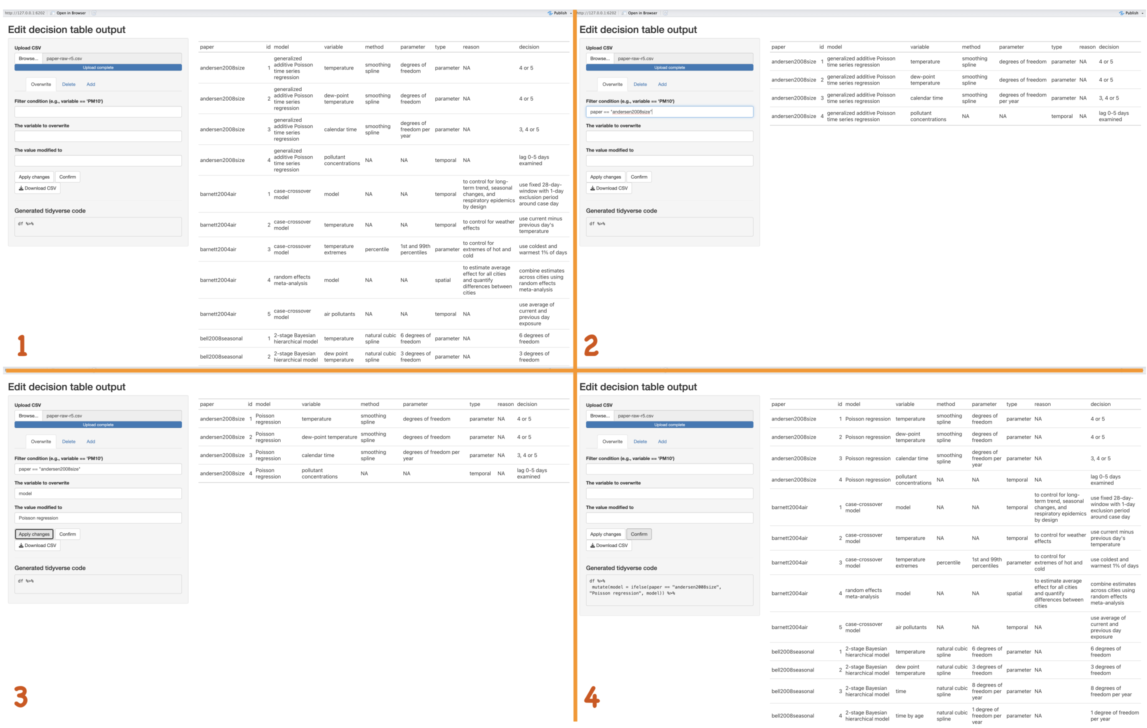

Fig. 1. The Shiny application interface for editting Large Language Model (LLM)-generated decisions (overwrite, delete, and add). (1) the default interface after loading the input CSV file. (2) The table view will update interactively upon the user-defined filter condition – expressed using `dplyr::filter()` syntax (e.g., `paper == anderson2008size`"), (3) The user edits the `model` column to “Poisson regression” and applies the change by clicking the Apply changes button. The table view updates to reflect the changes (4) After clicking the Confirm button, the corresponding `tidyverse` code is generated, and the table view returns to its original unfiltered view. The edited data can be downloaded by clicking the Download CSV button.

4 Calculating paper similarity

Once the decisions have been extracted and validated, this opens up a structured data for analyzing these information. For example, we can compare whether author's choices at different times changes, or across decisions varies at different regions. In this section, we present a method to calculate paper similarity based on the decisions shared in the paper pairs. The goal is to construct a distance metric based on similarity of the decision choice among papers that could be further used for clustering paper based on choices made by different authors in the literature. An overview of the method is illustrated in Figure 2.

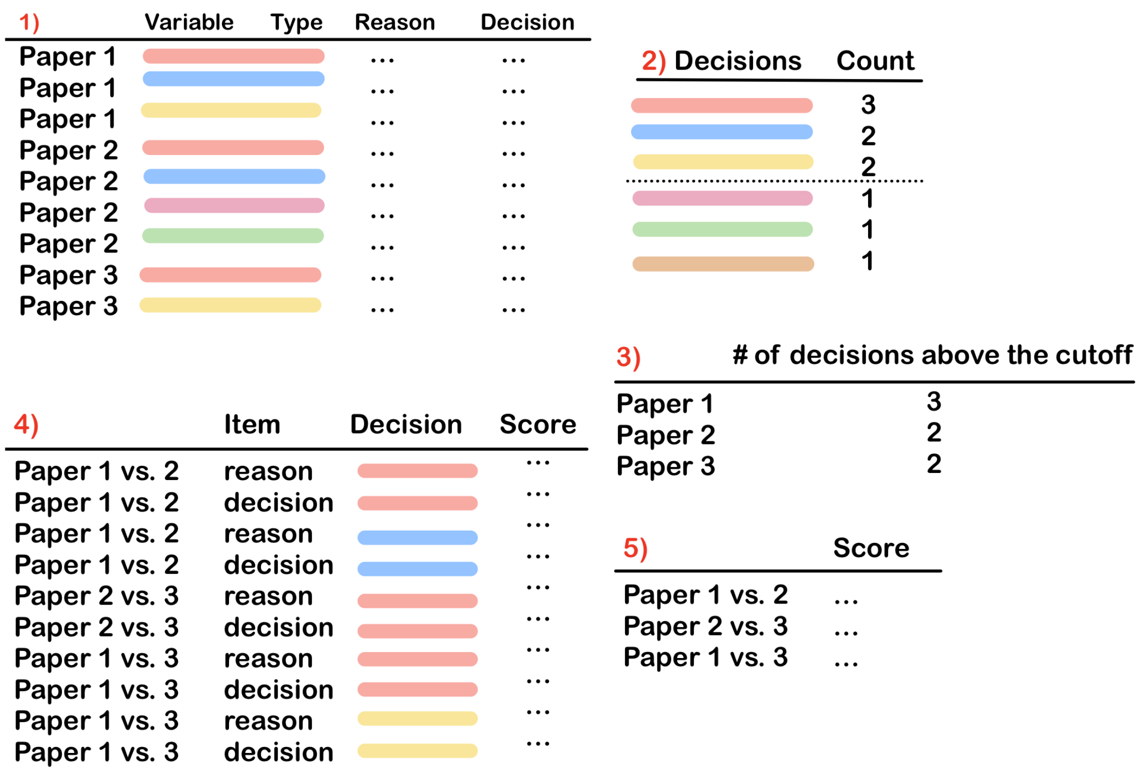

Fig. 2. Workflow for calculating paper similarity based on decision choices: (1) standardize variable names, (2) identify most frequent variable-type decisions across all papers, (3) identify papers with at least x identified decisions, (4) calculate decisions similarity score on the *decision* and *reason* fields using transformer language models, e.g. BERT, (5) calculate paper similarity score based on aggregating decision similarity scores.

- **TODO** some discussion on what it means by for two papers to be similar based on decisions.

The calculation of paper similarity is based on the similarity of decisions shared by each paper pair. A decision comparable in two papers are the ones that share the same variable and type, e.g. temperature and parameter (a decisions on the choosing the statistical method *parameter* for the *temperature* variable), or humidity and temporal (any *temporal* treatment, e.g. choice of lag value for the *humidity* variable). While many decisions share a similar variable, different authors may refer to them with slightly different names, such as "mean temperature" and "average temperature", hence variable names are first standardized to a common set of variable names. For example, "mean temperature" and "average temperature" are both standardized to "temperature". Notice that "dewpoint temperature" is standardized into "humidity" since it is a proxy of temperature to achieve a relative humidity (RH) of 100%. For literature with a common theme, there is usually a set of variables that shared by most papers and additional variables are justified in individual research. For our air pollution mortality modelling literature, we standardize the following variable names:

- **temperature**: "mean temperature", "average temperature", "temperature", "air temperature", "ambient temperature"
- **humidity**: "dewpoint temperature" and its hyphenated variants, relative humidity", "humidity"
- **PM**: "pollutant", "pollution", "particulate matter", "particulate", "PM10", "PM2.5"
- **time**: "date", "time", "trends", "trend"

Depending on the specific pairs, papers have varied number of decisions that can be compared and aggregated. While paper similarities can be computed for all paper pairs, using the similarity of one or two pair of decisions to represent paper similarity is less ideal. Hence, before calculating the text similarity of decisions, we also include two optional steps to identify and subset the most frequent decisions across papers, and to retain only papers that report more than a certain number of frequent decisions. Research questions in different fields may have different levels of homogeneity, depending on the maturity of the field and for air pollution mortality modelling, it is helpful to focus on decisions and papers that share a substantial number of decisions.

To assign numerical value for the similarity of reason, we use a transformer language model, such as BERT, to measure the semantic text similarity between the decision itself and its justification. The decision similarity is calculated by comparing the *decision* and *reason* fields of the decisions in each paper pair. To obtain paper similarity, we average the decision similarities across all decisions in each paper pair and other method can be customized for aggregation. The resulting paper similarity score can be used as a distance matrix to cluster papers based on their decision choices to understand the common practices in the investigated literature.

5 Results

From the 57 studies examining the effect of particulate matters (PM_{10} and $PM_{2.5}$) on mortality, we focus on the baseline model reported in each paper, excluding secondary models (e.g. lag-distributed models) and sensitivity analysis. We also exclude decisions on other pollutants, such as nitrogen dioxide (NO_2). This yields 273 decisions extracted using Gemini, averaging approximately 5 decisions per paper. Table 2 summarizes the number of edits made during the review process using the Shiny app. [details]

Table 3 summarizes the missingness of the decisions and reason. While most papers report their decision choices (e.g. use of five degree of freedom), 57% of decisions lack a stated rationale for the choice. Table 4 lists teh eight most frequently reported decision: parameter and temporal choice for time, PM, temperature, and humidity.

Table 2. tsdjflkajsldf.

Reason	Count
Recode for secondary LLM processing for standardization	42
Irrelevant decisions, e.g. other pollutants, sensitivity analysis	36
Decision captured not correct	10
Duplicates	9
General statements without specific decision, e.g. minimum of 1 df per year was required	6

Table 3. Missingness of decision and reason fields in the Gemini-extracted decisions. Most decisions report the choice (35.5 + 57.1 = 92%), but 57.1% lacks a stated reason.

	Decision	
Reason	Non-missing	Missing
Non-missing	97 (35.5%)	16 (5.9%)
Missing	156 (57.1%)	4 (1.5%)

Table 2. tsdjflkajsldf.

Reason	Count
Definition of variables, e.g. season	5
Total	108

Table 4. Count of variable-type decisions in the Gemini-extracted decisions. The most commonly reported decision are the parameter choices and temporal lags for for time, PM, temperature, and humidity.

Variable	Type	Count
PM	temporal	56
time	parameter	45
temperature	parameter	37
humidity	parameter	26
temperature	temporal	25
humidity	temporal	20
PM	parameter	8
mortality	temporal	4

Table 5 reports the parameter-related decisions captured in the literature. They refer to the number of knots or degree of freedom for spline methods (natural and smoothing spline) applied to variable time, humidity and temperature. For consistency, all values have been converted to a *per year* scale. The selection of knot for natural spline has less variation than the degree of freedom choices for smoothing spline. Choices for temperature and humidity tend to be close, given they are both weather related variables, while the choices for time are more varied inherently. This tabulation offers a reference set for potential options for future studies and help to identify anomalies and special treatment in practice. Notable example includes the use of 7.7 degree of freedom in Castillejos et al. [12], and highly flexible choices of 30 and 100 in Moolgavkar [34] and Moolgavkar [35], respectively. While most papers choice to report the smoothing parameter as a constant value, Schwartz [40] specifies it as a proportion of the data ("5% of the data" and "5% of the data").

For temporal decisions, after an initial review, we observed that decisions are still highly varied. The decisions can be divided into two groups: multi-day lags include expressions such as "6-day average", "3-d moving average", "mean of lags 0+1", and "cumulative lags, mean 0+1+2", and single-day lags include "lagged exposure up to 6 days", "lag days from 0 to 5" among others. To standardize these entries, we applied a secondary LLM process (claude-3-7-sonnet-latest) and converted them into a consistent format: `multi-day: lag [start]-[end]` and `single-day: lag [start], ... lag [end]`. Table 6 summarizes the temporal lag choices for PM, temperature, and humidity. Both single and multiple day lags are generally considered up to five days prior (lag 5). [TODO: check multi-day starts from one].

Table 5. Options captured for parameter choices for time, humidity, and temperature variables in the Gemini-extracted decisions. The choices for natural spline knots are generally less varied than the degree of freedom choices for smoothing spline. Choices for temperature and humidity tend to be close, given they are both weather related variables, while the choices for time are more varied inherently.

Method	Variable	Decision
natural spline	humidity	3, 4
natural spline	temperature	3, 4, 6
natural spline	time	1, 3, 4, 6, 7, 8, 12
smoothing spline	humidity	2, 3, 4, 6, 8, 50% of the data
smoothing spline	temperature	2, 3, 4, 6, 8, 50% of the data
smoothing spline	time	1, 3, 4, 5, 6, 7, 7.7, 8, 9, 10, 12, 30, 100, NA, 5% of the data

Table 6. Options captured for temporal lag choices for PM, temperature, and humidity variables in the Gemini-extracted decisions. Both single-day lags and multi-day average lags are commonly used, generally considering up to five days prior (lag 5).

Lag type	Variable	Decision
multi-day average	PM	0-1, 0-2, 0-3, 0-4, 0-5, 0-7, 1-0, 1-5
multi-day average	humidity	0-1, 0-2, 0-3, 0-5, 1-3, 1-5
multi-day average	temperature	0-1, 0-2, 0-3, 0-5, 0-7, 1-3
single-day lag	PM	lag 0, lag 1, lag 2, lag 3, lag 4, lag 5
single-day lag	humidity	lag 0, lag 1, lag 2, lag 3, lag 4, lag 5, lag 6, lag 7
single-day lag	temperature	lag 0, lag 1, lag 2, lag 3, lag 4, lag 5

For computing the decision similarity score, we include the first 6 most common variable-type decisions as suggested in Table 4. Figure 4 shows the clustering of the 50 papers based on the decision similarity scores. The dendrogram is generated using hierarchical clustering, and the labels are colored according to the most common smoothing method used in each paper. The clustering reveals three distinct groups of papers, which reflect the modelling strategies differ in the European (LOESS) and U.S. (...) studies [more on the APHENA].

6 Sensitivity analysis

In this section, we examine the reproducibility for using LLMs for text extraction tasks in Section 6.1, discrepancies between different LLM models: Gemini (`gemini-2.0-flash`) and Claude (`claude-3-7-sonnet-latest`) in Section 6.2, and the sensitivity of our paper similarity calculation pipeline to the choice of text model used for computing decision similarity scores in Section 6.3.

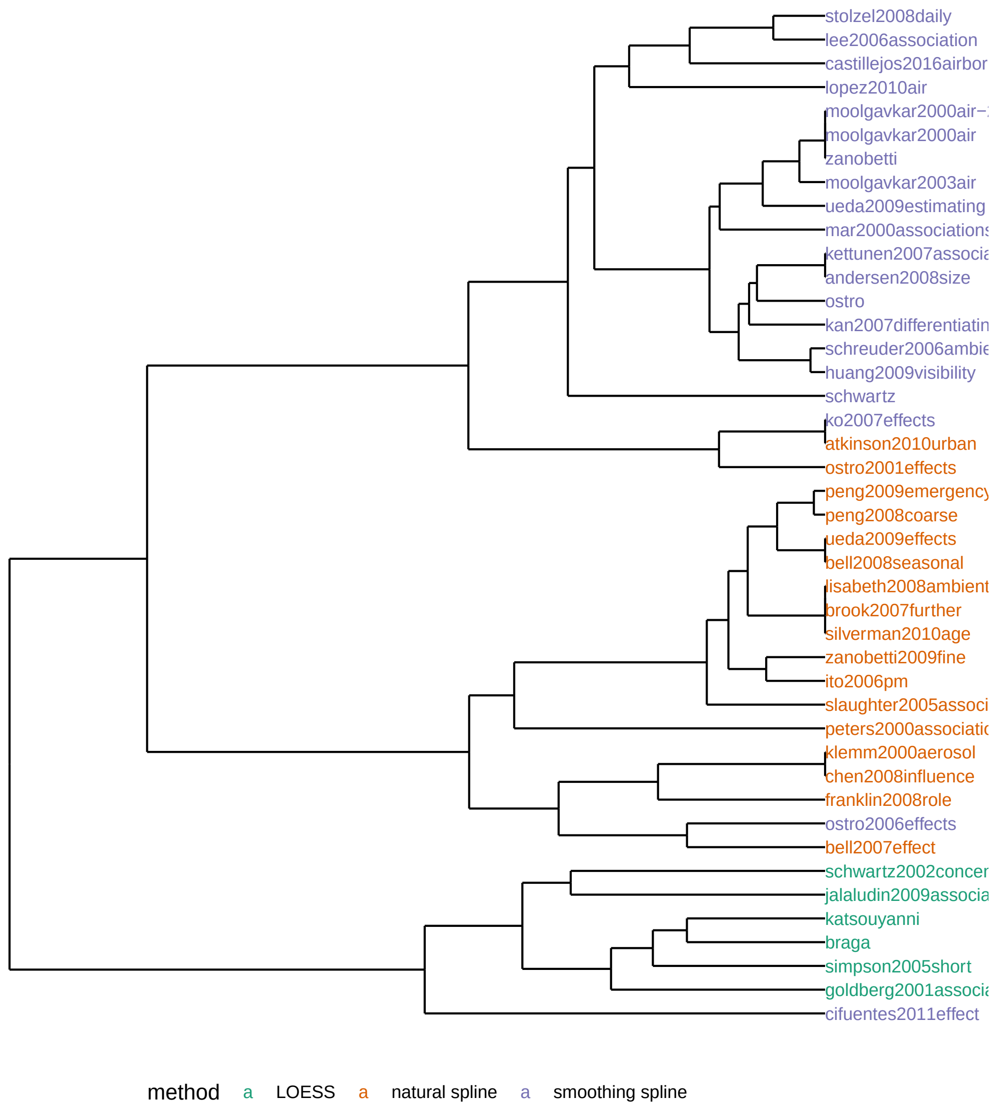

Fig. 3. The dendrogram (left) and multi-dimensional scaling (MDS) (right) based on paper similarity distance for 62 air pollution mortality modelling literature. The papers are colored by the most common smoothing method used. The MDS reveals the three distinct groups of papers. This grouping corresponds to the modelling strategies differ in the European and U.S. studies, documented in ALPHENA.

6.1 LLM reproducibility

For our text extraction task, we test the reproducibility of Gemini (`gemini-2.0-flash`) by repeating the text extraction task 5 times for each of the 62 papers. For each of the 31 papers, five runs yield $5 \times 4/2 = 10$ pairwise comparisons per field and including both the "reason" and "decision" fields results in a total of $31 \times 10 \times 2 = 620$ pairs. We exclude the

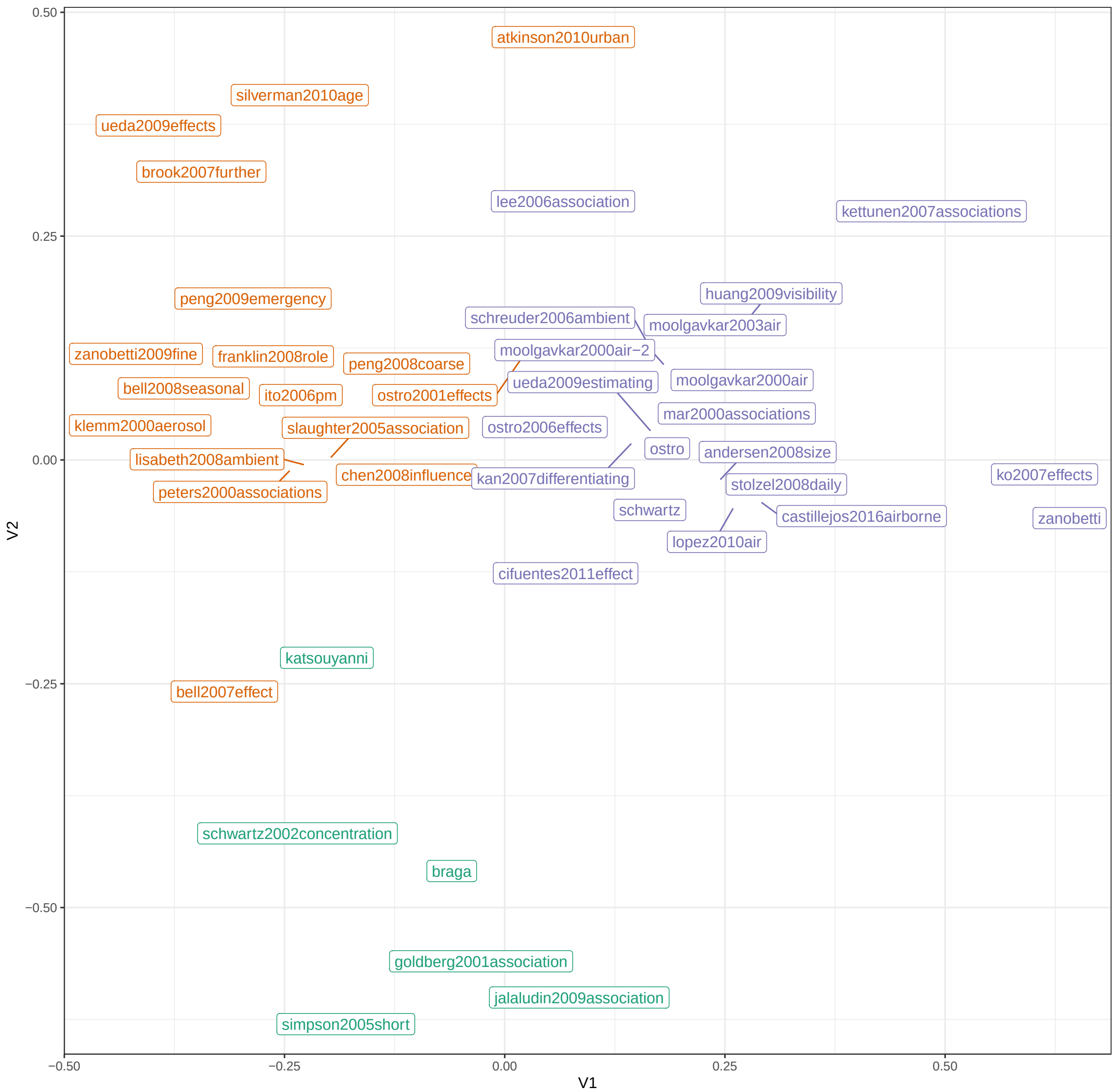

Fig. 4. The dendrogram (left) and multi-dimensional scaling (MDS) (right) based on paper similarity distance for 62 air pollution mortality modelling literature. The papers are colored by the most common smoothing method used. The MDS reveals the three distinct groups of papers. This grouping corresponds to the modelling strategies differ in the European and U.S. studies, documented in ALPHENA.

pairs that have different number of decisions since it would require manually align the decision to compare and this left us with 449 out of 620 (72%) pairwise comparisons. Table 7 shows an example of such comparison in Andersen et al. [3], where all the four reasons are identical among the two runs, hence a zero number of difference.

Table 7. An example of comparing the text extraction in decisions in Andersen 2008.

Variable	Run1	Run2
NCtot	6day average (lag 05)	6day average (lag 05)
calendar time	3 4 or 5 dfyear	3 4 or 5 dfyear
dew-point temperature	4 or 5 df	4 or 5 df
temperature	4 or 5 df	4 or 5 df

Table 8 summarizes the number of differences observed in each pairwise comparison. Among all comparisons, 80% produce the identical text in reason and decision. The discrepancies come from the following reasons:

- Gemini extracted different length for the same decision, e.g. in Kan et al. [25], some runs may extract "singleday lag models underestimate the cumulative effect of pollutants on mortality 2day moving average **of current and previous day concentrations** (lag=01)", while others extract "singleday lag models underestimate the cumulative effect of pollutants on mortality 2day moving average (lag=01)". Similarity, for decisions, some runs may yield "10 df for total mortality", while other runs yield "10 df". Similar extraction appears in Breitner et al. [9].
- Gemini fails to extract reasons in some runs but not others, e.g. in Burnett et al. [10], the first run generates NAs in the reasons, but the remaining four runs are identical. In Ueda et al. [44] and Castillejos et al. [12] , runs 1 and 5 fail to extract the reasons and produce the same incomplete version, whereas runs 2, 3, and 4 produce accurate versions with reasons populated.

Table 8. Number of differences in the reason and decision fields across Gemini runs for papers with consistent number of decisions across runs.

Num. of difference	Count	Proportion (%)
0	358	79.73
1	12	2.67
2	8	1.78
3	0	0.00
4	24	5.35
5	12	2.67
6	3	0.67
7	0	0.00
8	10	2.23
9	6	1.34
10	10	2.23
11	6	1.34
Total	449	100.00

6.2 LLM models

Reading text from PDF document requires Optical Character Recognition (OCR) to convert images into machine-readable text, which currently is only supported by Antropic Claude (`claude-3-7-sonnet-latest`) and Google Gemini (`gemini-2.0-flash`).

We compare the number of decisions extracted by Claude and Gemini across all 62 papers in **?@fig-claude-gemini**. Each point represents a paper, with the x- and y-axes showing the number of decisions extracted by Claude and Gemini, respectively. The dashed 1:1 line marks where both models extract the same number of decisions. Most points fall below this line, indicating that Claude extracts more decisions – often from data pre-processing or secondary data analysis steps requiring more manual validation – whereas Gemini focuses more on modelling choices relevant to our analysis. Some of these decisions captured by Claude are

- the definition of "cold day" and "hot day" indicators in Dockery et al. [16] ("defined at the 5th/ 95th percentile"),
- the choice to summarize NO_2, O_3, and SO_2 using a "24 hr average on variable" in Huang et al. [21], and
- the definition of black smoke and in Katsouyanni et al. [26] for secondary analysis ("restrict to days with BS concentrations below 150 $\mu g/m^2$").

Gemini sometimes also include irrelevant decisions, such as in Mar et al. [33], where secondary analysis choices like "0-4 lag days" for air pollution exposure variables (CO, EC, K_S, NO_2, O_3, OC, Pb, S, SO_2, TC, Zn) are captured. However, these cases are less frequent, resulting in outputs with less noise overall.

For both Claude and Gemini, we find they fail to link the general term "weather variables" to the specific weather variables. For example Gemini misses this link in Dockery et al. [16] and Burnett et al. [11], while Claude does so in Dockery et al. [16] and Katsouyanni et al. [26]. Although our prompt specified that some decisions may require linking information across sentences and paragraphs to identify the correct variable, this instruction doesn't appear to be applied consistently.

6.3 Text model

We have conducted sensitivity analysis on the text model for obtaining the decision similarity score from the Gemini outputs. The tested language models tested include

1) BERT by Google [15],
2) RoBERTa by Facebook AI [32], trained on a larger dataset (160GB v.s. BERT's 15GB),
3) XLNnet by Google Brain [47], and

two domain-trained BERT models:

4) sciBERT [4], trained on scientific literature, and
5) bioBERT [27], trained on PubMed and PMC data.

Figure 5 presents the distribution of the decision similarity (left) and paper similarity (right) for each text model. At decision level, the BERT model produces the widest variation across all five models, while the similarity scores from XLNet are all close to 1. These scores are not comparable across models since the difference of the underlying transformer architecture. However, the paper similarity scores from each model are comparable and **?@fig-text-mds** shows the multi-dimensional scaling (MDS) of the paper similarity scores from each text model: all showing a similar clustering pattern of the three main smoothing methods.

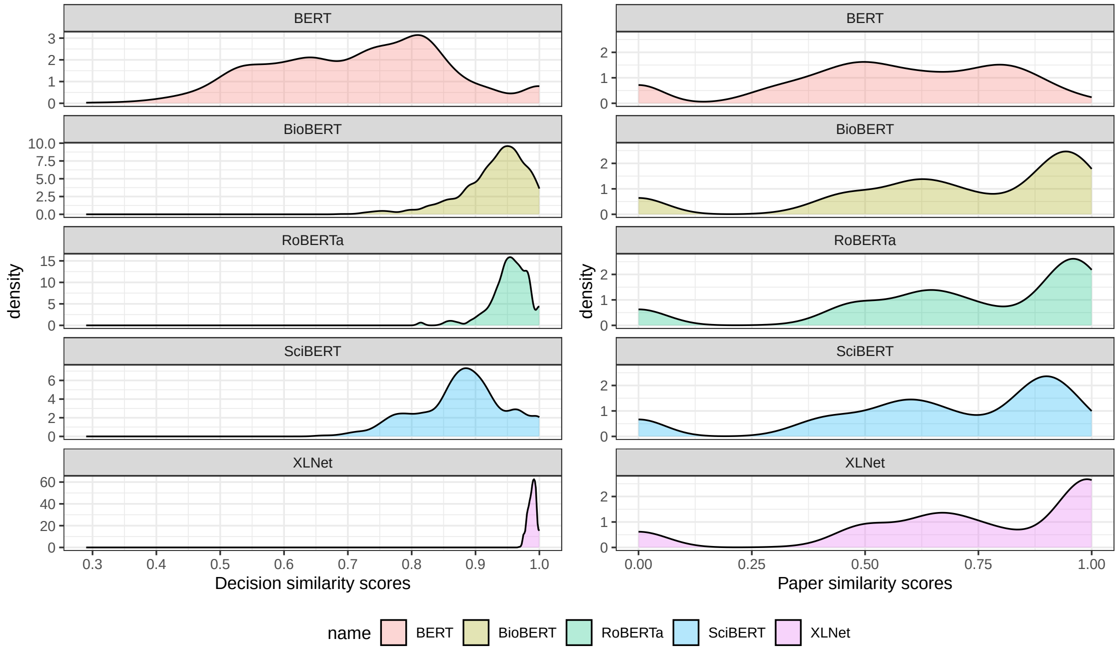

Fig. 5. Distribution of decision similarity (left) and paper similarity (right) scores for five different text models (BERT, BioBERT, RoBERTa, SciBERT, and XLNet). The default language model, BERT, produces the widest variation across the five models, while the similarity scores form XLNet are all close to 1. The model BioBERT, RoBERTa, and SciBERT yield decision similar scores mostly between 0.7 to 1.

7 Discussion

- Only prompting engineering is used to extract decisions from the literature. We expect that fine-tuning the model on statistical or domain-specific literature to yield more robust performance on the same document, though it would require substantially more training effort.
- people from the NYU-LMU workshop are interested to have code script attached as well because people can do one thing in the script but report another in the paper - it would be interesting to compare the paper and the script with some syntax extraction.
- Spatial decisions are generally not well captured because it often conducted uniformly as estimating the city individually to accommodate city heterogeneity. Some papers only consider a handful of cities, while in larger studies the individual city effects are then pooled together using random effect.
- Validation of the output: the nature of the task: Our task involve a reasoning component in that it requires casual reasoning to identify the decisions made by the authors, and its justification/ rationale, rather than purely summarizing the text through pattern-matching.
- the variation of parameters people use can help to identify parameter that needs a separate sensitivity analysis

References

[1] Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. 2014 ieee conference on visual analytics science and technology (vast). pages 173–182, 10 2014. doi: 10.1109/VAST.2014.7042493. URL https://ieeexplore.ieee.org/document/7042493.

[2] Sara Alspaugh, Nava Zokaei, Andrea Liu, Cindy Jin, and Marti A. Hearst. Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):22–31, 01 2019. doi: 10.1109/TVCG.2018.2865040. URL https://ieeexplore.ieee.org/document/8440815.

[3] Z. J. Andersen, P. Wahlin, O. Raaschou-Nielsen, M. Ketzel, T. Scheike, and S. Loft. Size distribution and total number concentration of ultrafine and accumulation mode particles and hospital admissions in children and the elderly in copenhagen, denmark. *Occupational and Environmental Medicine*, 65(7):458–466, 07 2008. doi: 10.1136/oem.2007.033290. URL https://oem.bmj.com/content/65/7/458. Publisher: BMJ Publishing Group Ltd Section: Original article PMID: 17989204.

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp). pages 3613–3618, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://www.aclweb.org/anthology/D19-1371.

[5] Steven Bethard and Dan Jurafsky. Cikm '10: International conference on information and knowledge management. pages 609–618, Toronto ON Canada, 10 2010. ACM. doi: 10.1145/1871437.1871517. URL https://dl.acm.org/doi/10.1145/1871437.1871517.

[6] Dorothy V. M. Bishop and Charles Hulme. When alternative analyses of the same data come to different conclusions: A tutorial using declaredesign with a worked real-world example. *Advances in Methods and Practices in Psychological Science*, 7(3):25152459241267904, 07 2024. doi: 10.1177/25152459241267904. URL https://doi.org/10.1177/25152459241267904. Publisher: SAGE Publications Inc.

[7] Graeme Blair, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. Declaring and diagnosing research designs. *American Political Science Review*, 113(3):838–859, 08 2019. doi: 10.1017/S0003055419000194. URL https://www.cambridge.org/core/product/identifier/S0003055419000194/type/journal_article.

[8] Rotem Botvinik-Nezer, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A. Mumford, R. Alison Adcock, Paolo Avesani, Blazej M. Baczkowski, Aahana Bajracharya, Leah Bakst, Sheryl Ball, Marco Barilari, Nadège Bault, Derek Beaton, Julia Beitner, Roland G. Benoit, Ruud M. W. J. Berkers, Jamil P. Bhanji, Bharat B. Biswal, Sebastian Bobadilla-Suarez, Tiago Bortolini, Katherine L. Bottenhorn, Alexander Bowring, Senne Braem, Hayley R. Brooks, Emily G. Brudner, Cristian B. Calderon, Julia A. Camilleri, Jaime J. Castrellon, Luca Cecchetti, Edna C. Cieslik, Zachary J. Cole, Olivier Collignon, Robert W. Cox, William A. Cunningham, Stefan Czoschke, Kamalaker Dadi, Charles P. Davis, Alberto De Luca, Mauricio R. Delgado, Lysia Demetriou, Jeffrey B. Dennison, Xin Di, Erin W. Dickie, Ekaterina Dobryakova, Claire L. Donnat, Juergen Dukart, Niall W. Duncan, Joke Durnez, Amr Eed, Simon B. Eickhoff, Andrew Erhart, Laura Fontanesi, G. Matthew Fricke, Shiguang Fu, Adriana Galván, Remi Gau, Sarah Genon, Tristan Glatard, Enrico Glerean, Jelle J. Goeman, Sergej A. E. Golowin, Carlos González-García, Krzysztof J. Gorgolewski, Cheryl L. Grady, Mikella A. Green, João F. Guassi Moreira, Olivia Guest, Shabnam Hakimi, J. Paul Hamilton, Roeland Hancock, Giacomo Handjaras, Bronson B. Harry, Colin Hawco, Peer Herholz, Gabrielle Herman, Stephan Heunis, Felix Hoffstaedter, Jeremy Hogeveen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacovella, Alexandru D. Iordan, Peder M. Isager, Ayse I. Isik, Andrew Jahn, Matthew R. Johnson, Tom Johnstone, Michael J. E. Joseph, Anthony C. Juliano, Joseph W. Kable, Michalis Kassinopoulos, Cemal Koba, Xiang-Zhen Kong, Timothy R. Koscik, Nuri Erkut Kucukboyaci, Brice A. Kuhl, Sebastian Kupek, Angela R. Laird, Claus Lamm, Robert Langner, Nina Lauharatanahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Andrea Leo, Elise Lesage, Flora Li, Monica Y. C. Li, Phui Cheng Lim, Evan N. Lintz, Schuyler W. Liphardt, Annabel B. Losecaat Vermeer, Bradley C. Love, Michael L. Mack, Norberto Malpica, Theo Marins, Camille Maumet, Kelsey McDonald, Joseph T. McGuire, Helena Melero, Adriana S. Méndez Leal, Benjamin Meyer, Kristin N. Meyer, Glad Mihai, Georgios D. Mitsis, Jorge Moll, Dylan M. Nielson, Gustav Nilsonne, Michael P. Notter, Emanuele Olivetti, Adrian I. Onicas, Paolo Papale, Kaustubh R. Patil, Jonathan E. Peelle, Alexandre Pérez, Doris Pischedda, Jean-Baptiste Poline, Yanina Prystauka, Shruti Ray, Patricia A. Reuter-Lorenz, Richard C. Reynolds, Emiliano Ricciardi, Jenny R. Rieck, Anais M. Rodriguez-Thompson, Anthony Romyn, Taylor Salo, Gregory R. Samanez-Larkin, Emilio Sanz-Morales, Margaret L. Schlichting, Douglas H. Schultz, Qiang Shen, Margaret A. Sheridan, Jennifer A. Silvers, Kenny Skagerlund, Alec Smith, David V. Smith, Peter Sokol-Hessner, Simon R. Steinkamp, Sarah M. Tashjian, Bertrand Thirion, John N. Thorp, Gustav Tinghög, Loreen Tisdall, Steven H. Tompson, Claudio Toro-Serey, Juan Jesus Torre Tresols, Leonardo Tozzi, Vuong Truong, Luca Turella, Anna E. van 't Veer, Tom Verguts, Jean M. Vettel, Sagana Vijayarajah, Khoi Vo, Matthew B. Wall, Wouter D. Weeda, Susanne Weis, David J. White, David Wisniewski, Alba Xifra-Porxas, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S. L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E. Nichols, Russell A. Poldrack, and Tom Schonberg. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810): 84–88, 06 2020. doi: 10.1038/s41586-020-2314-9. URL https://www.nature.com/articles/s41586-020-2314-9. Publisher: Nature Publishing Group.

[9] Susanne Breitner, Matthias Stölzel, Josef Cyrys, Mike Pitz, Gabriele Wölke, Wolfgang Kreyling, Helmut Küchenhoff, Joachim Heinrich, H.-Erich Wichmann, and Annette Peters. Short-term mortality rates during a decade of improved air quality in erfurt, germany. *Environmental Health Perspectives*, 117(3):448–454, 03 2009. doi: 10.1289/ehp.11711. URL https://ehp.niehs.nih.gov/doi/10.1289/ehp.11711. Publisher: Environmental Health Perspectives.

[10] Richard T. Burnett, Sabit Cakmak, Mark E. Raizenne, David Stieb, Renaud Vincent, Daniel Krewski, Jeffrey R. Brook, Owen Philips, and Haluk Ozkaynak. The association between ambient carbon monoxide levels and daily mortality in toronto, canada. *Journal of the Air & Waste Management Association*, 48(8):689–700, 08 1998. doi: 10.1080/10473289.1998.10463718. URL https://www.tandfonline.com/doi/full/10.1080/10473289.1998.10463718.

[11] Richard T. Burnett, Stieb ,Dave , Brook ,Jeffrey R. , Cakmak ,Sabit , Dales ,Robert , Raizenne ,Mark , Vincent ,Renaud , , and Tom Dann. Associations between short-term changes in nitrogen dioxide and mortality in canadian cities. *Archives of Environmental Health: An International Journal*, 59(5):228–236, 05 2004. doi: 10.3200/AEOH.59.5.228-236. URL https://doi.org/10.3200/AEOH.59.5.228-236. Publisher: Routledge _eprint: https://doi.org/10.3200/AEOH.59.5.228-236 PMID: 16201668.

[12] Margarita Castillejos, Borja-Aburto ,Victor H. , Dockery ,Douglas W. , Gold ,Diane R. , , and Dana. Loomis. Airborne coarse particles and mortality. *Inhalation Toxicology*, 12(sup1):61–72, 01 2000. doi: 10.1080/0895-8378.1987.11463182. URL https://doi.org/10.1080/0895-8378.1987.11463182. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0895-8378.1987.11463182.

[13] Chaomei Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, 2006. doi: 10.1002/asi.20317. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20317. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20317.

[14] J. K. Chou and C. K. Yang. Papervis: Literature review made easy. *Computer Graphics Forum*, 30(3):721–730, 2011. doi: 10.1111/j.1467-8659.2011.01921.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2011.01921.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2011.01921.x.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Naacl-hlt 2019. page 4171–4186, Minneapolis, Minnesota, 06 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

[16] Douglas W. Dockery, Joel Schwartz, and John D. Spengler. Air pollution and daily mortality: Associations with particulates and acid aerosols. *Environmental Research*, 59(2):362–373, 12 1992. doi: 10.1016/S0013-9351(05)80042-8. URL https://www.sciencedirect.com/science/article/pii/S0013935105800428.

[17] Marian Dörk, Nathalie Henry Riche, Gonzalo Ramos, and Susan Dumais. Pivotpaths: Strolling through faceted information spaces. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2709–2718, 12 2012. doi: 10.1109/TVCG.2012.252. URL https://ieeexplore.ieee.org/document/6327277.

[18] Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6):460–465, 12 2014. URL https://www.proquest.com/docview/1616141998/abstract/5E050DCE82414037PQ/1. Num Pages: 6 Place: Research Triangle Park, United States Publisher: Sigma XI-The Scientific Research Society.

[19] Elliot Gould, Hannah S. Fraser, Timothy H. Parker, Shinichi Nakagawa, Simon C. Griffith, Peter A. Vesk, Fiona Fidler, Daniel G. Hamilton, Robin N. Abbey-Lee, Jessica K. Abbott, Luis A. Aguirre, Carles Alcaraz, Irith Aloni, Drew Altschul, Kunal Arekar, Jeff W. Atkins, Joe Atkinson, Christopher M. Baker, Meghan Barrett, Kristian Bell, Suleiman Kehinde Bello, Iván Beltrán, Bernd J. Berauer, Michael Grant Bertram, Peter D. Billman, Charlie K. Blake, Shannon Blake, Louis Bliard, Andrea Bonisoli-Alquati, Timothée Bonnet, Camille Nina Marion Bordes, Aneesh P. H. Bose, Thomas Botterill-James, Melissa Anna Boyd, Sarah A. Boyle, Tom Bradfer-Lawrence, Jennifer Bradham, Jack A. Brand, Martin I. Brengdahl, Martin Bulla, Luc Bussière, Ettore Camerlenghi, Sara E. Campbell, Leonardo L. F. Campos, Anthony Caravaggi, Pedro Cardoso, Charles J. W. Carroll, Therese A. Catanach, Xuan Chen, Heung Ying Janet Chik, Emily Sarah Choy, Alec Philip Christie, Angela Chuang, Amanda J. Chunco, Bethany L. Clark, Andrea Contina, Garth A. Covernton, Murray P. Cox, Kimberly A. Cressman, Marco Crotti, Connor Davidson Crouch, Pietro B. D'Amelio, Alexandra Allison de Sousa, Timm Fabian Döbert, Ralph Dobler, Adam J. Dobson, Tim S. Doherty, Szymon Marian Drobniak, Alexandra Grace Duffy, Alison B. Duncan, Robert P. Dunn, Jamie Dunning, Trishna Dutta, Luke Eberhart-Hertel, Jared Alan Elmore, Mahmoud Medhat Elsherif, Holly M. English, David C. Ensminger, Ulrich Rainer Ernst, Stephen M. Ferguson, Esteban Fernandez-Juricic, Thalita Ferreira-Arruda, John Fieberg, Elizabeth A. Finch, Evan A. Fiorenza, David N. Fisher, Amélie Fontaine, Wolfgang Forstmeier, Yoan Fourcade, Graham S. Frank, Cathryn A. Freund, Eduardo Fuentes-Lillo, Sara L. Gandy, Dustin G. Gannon, Ana I. García-Cervigón, Alexis C. Garretson, Xuezhen Ge, William L. Geary, Charly Géron, Marc Gilles, Antje Girndt, Daniel Gliksman, Harrison B. Goldspiel, Dylan G. E. Gomes, Megan Kate Good, Sarah C. Goslee, J. Stephen Gosnell, Eliza M. Grames, Paolo Gratton, Nicholas M. Grebe, Skye M. Greenler, Maaike Griffioen, Daniel M. Griffith, Frances J. Griffith, Jake J. Grossman, Ali Güncan, Stef Haesen, James G. Hagan, Heather A. Hager, Jonathan Philo Harris, Natasha Dean Harrison, Sarah Syedia Hasnain, Justin Chase Havird, Andrew J. Heaton, María Laura Herrera-Chaustre, Tanner J. Howard, Bin-Yan Hsu, Fabiola Iannarilli, Esperanza C. Iranzo, Erik N. K. Iverson, Saheed Olaide Jimoh, Douglas H. Johnson, Martin Johnsson, Jesse Jorna, Tommaso Jucker, Martin Jung, Ineta Kačergytė, Oliver Kaltz, Alison Ke, Clint D. Kelly, Katharine Keogan, Friedrich Wolfgang Keppeler, Alexander K. Killion, Dongmin Kim, David P. Kochan, Peter Korsten, Shan Kothari, Jonas Kuppler, Jillian M. Kusch, Malgorzata Lagisz, Kristen Marianne Lalla, Daniel J. Larkin, Courtney L. Larson, Katherine S. Lauck, M. Elise Lauterbur, Alan Law, Don-Jean Léandri-Breton, Jonas J. Lembrechts, Kiara L'Herpiniere, Eva J. P. Lievens, Daniela Oliveira de Lima, Shane Lindsay, Martin Luquet, Ross MacLeod, Kirsty H. Macphie, Kit Magellan, Magdalena M. Mair, Lisa E. Malm, Stefano Mammola, Caitlin P. Mandeville, Michael Manhart, Laura Milena Manrique-Garzon, Elina Mäntylä, Philippe Marchand, Benjamin Michael Marshall, Charles A. Martin, Dominic Andreas Martin, Jake Mitchell Martin, April Robin Martinig, Erin S. McCallum, Mark McCauley, Sabrina M. McNew, Scott J. Meiners, Thomas Merkling, Marcus Michelangeli, Maria Moiron, Bruno Moreira, Jennifer Mortensen, Benjamin Mos, Taofeek Olatunbosun Muraina, Penelope Wrenn Murphy, Luca Nelli, Petri Niemelä, Josh Nightingale, Gustav Nilsonne, Sergio Nolazco, Sabine S. Nooten, Jessie Lanterman Novotny, Agnes Birgitta Olin, Chris L. Organ, Kate L. Ostevik, Facundo Xavier Palacio, Matthieu Paquet, Darren James Parker, David J. Pascall, Valerie J. Pasquarella, John Harold Paterson, Ana Payo-Payo, Karen Marie Pedersen, Grégoire Perez, Kayla I. Perry, Patrice Pottier, Michael J. Proulx, Raphaël Proulx, Jessica L. Pruett, Veronarindra Ramananjato, Finaritra Tolotra Randimbiarison, Onja H. Razafindratsima, Diana J. Rennison, Federico Riva, Sepand Riyahi, Michael James Roast, Felipe Pereira Rocha, Dominique G. Roche, Cristian Román-Palacios, Michael S. Rosenberg, Jessica Ross, Freya E. Rowland, Deusdedith Rugemalila, Avery L. Russell, Suvi Ruuskanen, Patrick Saccone, Asaf Sadeh, Stephen M. Salazar, Kris Sales, Pablo Salmón, Alfredo Sánchez-Tójar, Leticia Pereira Santos, Francesca Santostefano, Hayden T. Schilling, Marcus Schmidt, Tim Schmoll, Adam C. Schneider, Allie E. Schrock, Julia Schroeder, Nicolas Schtickzelle, Nick L. Schultz, Drew A. Scott, Michael Peter Scroggie, Julie Teresa Shapiro, Nitika Sharma, Caroline L. Shearer, Diego Simón, Michael I.

Sitvarin, Fabrício Luiz Skupien, Heather Lea Slinn, Grania Polly Smith, Jeremy A. Smith, Rahel Sollmann, Kaitlin Stack Whitney, Shannon Michael Still, Erica F. Stuber, Guy F. Sutton, Ben Swallow, Conor Claverie Taff, Elina Takola, Andrew J. Tanentzap, Rocío Tarjuelo, Richard J. Telford, Christopher J. Thawley, Hugo Thierry, Jacqueline Thomson, Svenja Tidau, Emily M. Tompkins, Claire Marie Tortorelli, Andrew Trlica, Biz R. Turnell, Lara Urban, Stijn Van de Vondel, Jessica Eva Megan van der Wal, Jens Van Eeckhoven, Francis van Oordt, K. Michelle Vanderwel, Mark C. Vanderwel, Karen J. Vanderwolf, Juliana Vélez, Diana Carolina Vergara-Florez, Brian C. Verrelli, Marcus Vinícius Vieira, Nora Villamil, Valerio Vitali, Julien Vollering, Jeffrey Walker, Xanthe J. Walker, Jonathan A. Walter, Pawel Waryszak, Ryan J. Weaver, Ronja E. M. Wedegärtner, Daniel L. Weller, and Shannon Whelan. Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology. *BMC Biology*, 23(1):35, 02 2025. doi: 10.1186/s12915-024-02101-x. URL https://doi.org/10.1186/s12915-024-02101-x.

[20] Florian Heimerl, Qi Han, Steffen Koch, and Thomas Ertl. Citerivers: Visual analytics of citation patterns. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):190–199, 01 2016. doi: 10.1109/TVCG.2015.2467621. URL https://ieeexplore.ieee.org/document/7192685/authors.

[21] Wei Huang, Jianguo Tan, Haidong Kan, Ni Zhao, Weimin Song, Guixiang Song, Guohai Chen, Lili Jiang, Cheng Jiang, Renjie Chen, and Bingheng Chen. Visibility, air quality and daily mortality in shanghai, china. *Science of The Total Environment*, 407(10):3295–3300, 05 2009. doi: 10.1016/j.scitotenv.2009.02.019. URL https://linkinghub.elsevier.com/retrieve/pii/S004896970900165X.

[22] Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, 59(3):944–960, 2021. doi: 10.1111/ecin.12992. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecin.12992. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecin.12992.

[23] Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Jian Chen, and Torsten Möller. Visualization as seen through its research paper keywords. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):771–780, 01 2017. doi: 10.1109/TVCG.2016.2598827. URL https://ieeexplore.ieee.org/document/7539364.

[24] Alex Kale, Matthew Kay, and Jessica Hullman. Decision-making under uncertainty in research synthesis: Designing for the garden of forking paths. CHI '19, page 1–14, New York, NY, USA, 05 2019. Association for Computing Machinery. doi: 10.1145/3290605.3300432. URL https://dl.acm.org/doi/10.1145/3290605.3300432.

[25] Haidong Kan, Stephanie J. London, Guohai Chen, Yunhui Zhang, Guixiang Song, Naiqing Zhao, Lili Jiang, and Bingheng Chen. Differentiating the effects of fine and coarse particles on daily mortality in shanghai, china. *Environment International*, 33(3):376–384, 04 2007. doi: 10.1016/j.envint.2006.12.001. URL https://www.sciencedirect.com/science/article/pii/S0160412006002108.

[26] Klea Katsouyanni, Giota Touloumi, Evangelia Samoli, Alexandros Gryparis, Alain Le Tertre, Yannis Monopolis, Giuseppe Rossi, Denis Zmirou, Ferran Ballester, Azedine Boumghar, Hugh Ross Anderson, Bogdan Wojtyniak, Anna Paldy, Rony Braunstein, Juha Pekkanen, Christian Schindler, and Joel Schwartz. Confounding and effect modification in the short-term effects of ambient particles on total mortality: Results from 29 european cities within the aphea2 project. *Epidemiology*, 12(5):521, 09 2001. URL https://journals.lww.com/epidem/fulltext/2001/09000/confounding_and_effect_modification_in_the.11.aspx.

[27] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 02 2020. doi: 10.1093/bioinformatics/btz682. URL https://academic.oup.com/bioinformatics/article/36/4/1234/5566506.

[28] S. L. Lee, W. H. S. Wong, and Y. L. Lau. Association between air pollution and asthma admission among children in hong kong. *Clinical & Experimental Allergy*, 36(9):1138–1146, 2006. doi: 10.1111/j.1365-2222.2006.02555.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2222.2006.02555.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2222.2006.02555.x.

[29] Jiali Liu, Nadia Boukhelifa, and James R. Eagan. Understanding the Role of Alternatives in Data Analysis Practices. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):66–76, January 2020. ISSN 1941-0506. doi: 10.1109/TVCG.2019.2934593. URL https://ieeexplore.ieee.org/document/8805460/.

[30] Yang Liu, Tim Althoff, and Jeffrey Heer. Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end data analysis. CHI '20, page 1–14, New York, NY, USA, 04 2020. Association for Computing Machinery. doi: 10.1145/3313831.3376533. URL https://dl.acm.org/doi/10.1145/3313831.3376533.

[31] Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. Boba: Authoring and visualizing multiverse analyses. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1753–1763, 02 2021. doi: 10.1109/TVCG.2020.3028985. URL https://ieeexplore.ieee.org/document/9216579/.

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. doi: 10.48550/arXiv.1907.11692.

[33] T F Mar, G A Norris, J Q Koenig, and T V Larson. Associations between air pollution and mortality in phoenix, 1995-1997. *Environmental Health Perspectives*, 108(4):347–353, 04 2000. doi: 10.1289/ehp.00108347. URL https://ehp.niehs.nih.gov/doi/abs/10.1289/ehp.00108347. Publisher: Environmental Health Perspectives.

[34] Suresh H. Moolgavkar. Air pollution and hospital admissions for diseases of the circulatory system in three u.s. metropolitan areas. *Journal of the Air & Waste Management Association*, 50(7):1199–1206, 07 2000. doi: 10.1080/10473289.2000.10464162. URL https://doi.org/10.1080/10473289.2000.10464162. Publisher: Taylor & Francis.

[35] Suresh H. Moolgavkar. Air pollution and daily mortality in two u.s. counties: Season-specific analyses and exposure-response relationships. *Inhalation Toxicology*, 15(9):877–907, 01 2003. doi: 10.1080/08958370390215767. URL https://doi.org/10.1080/08958370390215767. Publisher: Taylor & Francis.

[36] Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. Vitality: Promoting serendipitous discovery of academic literature with transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):486–496, 01 2022. doi: 10.1109/TVCG.2021.3114820. URL https://ieeexplore.ieee.org/document/9552447/.

[37] Bart Ostro, Rachel Broadwin, Shelley Green, Wen-Ying Feng, and Michael Lipsett. Fine particulate air pollution and mortality in nine california counties: Results from calfine. *Environmental Health Perspectives*, 114(1):29–33, 01 2006. doi: 10.1289/ehp.8335. URL https://ehp.niehs.nih.gov/doi/10.1289/ehp.8335. Publisher: Environmental Health Perspectives.

[38] Abhraneel Sarma, Alex Kale, Michael Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. multiverse: Multiplexing alternative data analyses in r notebooks (version 0.6.2). *OSF Preprints*, 2021. URL https://github.com/MUCollective/multiverse.

[39] Marko Sarstedt, Susanne J. Adler, Christian M. Ringle, Gyeongcheol Cho, Adamantios Diamantopoulos, Heungsun Hwang, and Benjamin D. Liengaard. Same model, same data, but different outcomes: Evaluating the impact of method choices in structural equation modeling. *Journal of Product Innovation Management*, 41(6):1100–1117, 2024. doi: 10.1111/jpim.12738. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jpim.12738. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jpim.12738.

[40] Joel Schwartz. The distributed lag between air pollution and daily deaths. *Epidemiology*, 11(3):320–326, 2000. URL https://www.jstor.org/stable/3703220. Publisher: Lippincott Williams & Wilkins.

[41] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 09 2018. doi: 10.1177/2515245917747646. URL https://doi.org/10.1177/2515245917747646. Publisher: SAGE Publications Inc.

[42] Jan Simson, Fiona Draxler, Samuel Mehr, and Christoph Kern. Preventing harmful data practices by using participatory input to navigate the machine learning multiverse. CHI '25, page 1–30, New York, NY, USA, 04 2025. Association for Computing Machinery. doi: 10.1145/3706598.3713482. URL https://dl.acm.org/doi/10.1145/3706598.3713482.

[43] Imad Tbahriti, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the medline digital library. *International Journal of Medical Informatics*, 75(6):488–495, 06 2006. doi: 10.1016/j.ijmedinf.2005.06.007. URL https://www.sciencedirect.com/science/article/pii/S1386505605000894.

[44] Kayo Ueda, Nitta ,Hiroshi , Ono ,Masaji , , and Ayano Takeuchi. Estimating mortality effects of fine particulate matter in japan: A comparison of time-series and case-crossover analyses. *Journal of the Air & Waste Management Association*, 59(10):1212–1218, 10 2009. doi: 10.3155/1047-3289.59.10.1212. URL https://doi.org/10.3155/1047-3289.59.10.1212. Publisher: Taylor & Francis _eprint: https://doi.org/10.3155/1047-3289.59.10.1212.

[45] Hadley Wickham. Tidy data. *Journal of Statistical Software*, 59:1–23, 09 2014. doi: 10.18637/jss.v059.i10. URL https://doi.org/10.18637/jss.v059.i10.

[46] Hadley Wickham, Joe Cheng, and Aaron Jacobs. *ellmer: Chat with Large Language Models*, 2025. URL https://CRAN.R-project.org/package=ellmer. R package version 0.1.1.

[47] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. doi: 10.48550/arXiv.1906.08237.