

1 **The Name of the Title Is Hope**

2
3 ANONYMOUS AUTHOR(S)

4
5 bla blabla

6
7 CCS Concepts: • **Applied computing** → *Document analysis; Environmental sciences;* • **Human-centered computing** → **HCI**
8 **theory, concepts and models.**

9 Additional Key Words and Phrases: Large language models

10 **ACM Reference Format:**

11 Anonymous Author(s). 2025. The Name of the Title Is Hope. In *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI'26)*. ACM, New York, NY, USA, 21 pages. <https://doi.org/XXXXXX.XXXXXXX>

- 12
13
14
15 • Something about “analysis review” - Roger thinks it’s a better to have a new word for this.
16 • provide a baseline understand - place to start
17 • demonstrate - analytically homogeneous - the table won’t look like that

18
19 **1 Introduction**

20
21 Decisions are everywhere in data analysis, from the initial data collection, data pre-processing to the modelling
22 choices. These decisions will impact the final output of the data analysis, which may lead to different conclusions
23 and policy recommendations. When such flexibility can be misused—through practices such as p-hacking, selective
24 reporting, or unjustified analytical adjustments—it can inflate effect sizes or produce misleading results that meet
25 conventional thresholds for statistical significance. They have been demonstrated through many-analysts experiments,
26 where independent teams analyzing the same dataset to answer a pre-defined research question often arrive at markedly
27 different conclusions. These practices not only compromise the validity of individual studies but also threaten the
28 broader credibility of statistical analysis and scientific research as a whole.

29
30 Multiple recommendations have been proposed to improve data analysis practices, such as pre-registration and
31 multiverse analysis. Bayesian methods also offer a different paradigm to p-value driven inference for interpreting
32 statistical evidence. Most empirical studies of data analysis practices focus on specially designed and simplified analysis
33 scenarios. While informative, these setups may not adequately capture the complexity of the data analysis with
34 significant policy implications. [In practice, studying the data analysis decisions with actual applications is challenging.]
35 Analysts may no longer be available for interviews due to job changes, and even when they are, recalling the full set
36 of decisions and thinking process made during the analysis is often infeasible. Moreover, only until the last decades,
37 analysis scripts and reproducible materials were not commonly required by journals for publishing. [As a result, it
38 remains challenging to study how analytical decisions are made.]

39
40 In this work, we focus on a specific class of air pollution modelling studies that estimate the effect size of particulate
41 matter (PM2.5 or PM10) on mortality, typically using Poisson regression or generalized additive models (GAMs).

42 Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not
43 made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components
44 of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to
45 redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

46 © 2025 ACM.

47 Manuscript submitted to ACM

48
49 Manuscript submitted to ACM

53 While individual modelling choices vary, these studies often share a common structure: they adjust for meteorological
 54 covariates such as temperature and humidity, apply temporal or spatial treatments, like including lagged variables and
 55 may estimate the effect by city or region before combining results. Because these studies investigate similar scientific
 56 questions using a shared modelling framework, they form a natural many-analyst setting. This allows us to examine, in
 57 a real-world context, the range of analytical decisions made by different researchers addressing the same underlying
 58 question.
 59

60 In this work, we develop a structured tabular format to record the analytical decisions made by researchers in the air
 61 pollution modelling literature. Using large language models (LLMs), we automate the extraction of these decisions from
 62 published papers. This allows us to treat decisions as data – allowing us to track them over time, compare methodology
 63 across papers, and query commonly used approaches. We further introduce a workflow to cluster studies based on
 64 decision similarity, revealing three distinct groups of papers that reflect the modelling strategies differ in the European
 65 and U.S. studies, which offers a new way to visualize the field in the air pollution mortality modelling.
 66

67 The rest of the paper is organized as follows. In Section 2, we review the background on data analysis decisions.
 68 Section 3 describes the data structure for recording decisions, the use of large language models to process research
 69 papers, and the validation of LLM outputs. In Section 4, we present the method for calculating paper similarity based
 70 on decision similarities. Section 5 reports the finding of our analysis, including the clustering of papers according to
 71 similarity scores and sensitivity analyses related to LLM providers, prompt engineering, and LLM parameters. Finally,
 72 Section 6 discusses the implications of our study.
 73

74 2 Background

75 2.1 Decisions in data analysis

76 **Question** Is “decision” going to be confusing with “decision-making” in decision theory

77 A data analysis is a process of making choices at each step, from the initial data collection to model specification, and
 78 post-processing. Each decision represents a branching point where analysts choose a specific path to follow, and the
 79 vast number of possible choices analysts can take forms what Gelman and Loken [18] describe as the “garden of forking
 80 paths”. While researchers may hope their inferential results are robust to the specific path taken through the garden,
 81 in practice, different choices can lead to substantially different conclusions. This has been empirically demonstrated
 82 through “many analyst experiments”, where independent research groups analyze the same dataset to the same answer
 83 using their chosen analytic approach. A classic example is Silberzahn et al. [36], where researchers reported an odds
 84 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
 85 variability has been observed in structural equation modeling [35], applied microeconomics [22], neuroimaging [8], and
 86 ecology and evolutionary biology [19]. Many studies have been conducted on a relatively smaller scale to interviews of
 87 analysts and researchers on data analysis practices [2, 24, 28], visualization of the decision process through the analytic
 88 decision graphics (ADG) [29]. Recently, Simson et al. [37] describes a participatory approach to decisions choices in
 89 fairness ML algorithms.
 90

91 Software tools have also developed to incorporate potential alternatives in the analysis workflow, including the
 92 DeclareDesign package [7] and the multiverse package [34]. The DeclareDesign package [7] introduces the MIDA
 93 framework for researchers to declare, diagnose, and redesign their analyses to produce a distribution of the statistic of
 94 interest, which has been applied in the randomized controlled trial study [6]. The multiverse package [34] provides
 95

105 a framework for researchers to systematically explore how different choices affect results and to report the range of
106 plausible outcomes that arise from alternative analytic paths.
107

108 2.2 Visualization on scientific literature

111 Much of the work on IEEE visualizing scientific literature focuses on helping researchers stay aware of relevant
112 publications, given the rapidly growing volume of scientific output and the difficulty of navigating it. Systems have
113 been developed to support the discovery of relevant papers, where relevance is typically determined by keywords [23],
114 citation information (e.g. citation list, co-citation) [13], or combinations with other relevant paper metadata (e.g. author,
115 title) [5, 14, 17, 20]. More recent approaches incorporate text-based information from the abstract or sections of the
116 paper to [obtain a better similar metric]. This includes using topic modelling [1], argumentation-based information
117 retrieval [38], and text embedding [32]. While these metadata information and high level text-based information are
118 valuable for discovering relevant papers, for data analysis, researchers need tools that help them *make sense* of the
119 literature rather than simply *finding* it. Capturing the decisions and reasoning expressed during analyses within a
120 similar theme can reveal common practices in the field and guide decisions choices in new applications. With recent
121 advances in Large Language Models (LLMs), it has become possible to automatically extract structured information from
122 unstructured text through prompting. This allows scientific literature to be clustered and visualized using information
123 about the underlying decisions and reasoning made during analysis, providing a basis for studying analysts' decision
124 choices.
125

126 2.3 Air pollution mortality modelling

127 3 Extracting decisions from data analysis

128 3.1 Decisions in data analysis

129 Decisions occur throughout the entire data analysis process – from the selection of variables and data source, to
130 pre-processing steps to prepare the data for modelling, to the model specification and variable inclusion. In this work,
131 we focus specifically on modelling decisions in the air pollution mortality modelling literature. These include the
132 choice of modelling approach, covariate inclusion and smoothing, and specifications of spatial and temporal structure.
133 Consider the following excerpt from Ostro et al. [33]:

134 Based on previous findings reported in the literature (e.g., Samet et al. 2000), the basic model included a
135 smoothing spline for time with 7 degrees of freedom (df) per year of data. This number of degrees of
136 freedom controls well for seasonal patterns in mortality and reduces and often eliminates autocorrelation.
137

138 This sentence encode the following components of a decision:

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

145 The decision above is regarding a certain parameter in the statistical method, we categorize this as a “parameter”
146 type decisions. Other types of decisions - such as spatial modelling structure or the inclusion of temporal lags - may
147

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 [40], 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. [33]:

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. [33]:

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.

209 3. **Some decisions may be omitted because they are data-driven.** For instance, Katsouyanni et al. [26] states:

210 The inclusion of lagged weather variables and the choice of smoothing parameters for all of the weather
211 variables were done by minimizing Akaike's information criterion.

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

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

222 223 224 225 3.2 Automatic reading of literature with LLMs

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

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

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

233 Text recognition from PDF document relies on Optical Character Recognition (OCR) to convert scanned images into
234 machine-readable text – capability currently offered by Anthropic Claude and Google Gemini. We instruct the LLM
235 to generate a markdown file containing a JSON block that records extracted decisions, which can then be read into
236 statistical software for further analysis. The exact prompt feed to the LLM is provided in the Appendix. The `ellmer`
237 package [41] in R is used to connect to the Gemini and Claude API, providing the PDF attachment and the prompt in a
238 markdown file as inputs.

243 244 245 246 3.3 Review the LLM output

- 247 • **TODO** something about result validation of LLM output

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

261 Edit decision table output
262 Uploaded CSV
263 Filter condition (e.g., variable == 'PM10')
264 The variable to overwrite
265 The value modified to
266 Apply changes Confirm
267 Generated tidyverse code
of NA

268 1
269 2
270 3
271 4
272 5
273 Edit decision table output
274 Uploaded CSV
275 Filter condition (e.g., variable == 'PM10')
276 The variable to overwrite
277 The value modified to
278 Poisson regression
279 Generated tidyverse code
of NA

280 3
281 4
282 5
283 6
284 7
285 8
286 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.

294
295
296
297
298
299
300
301
302 4 Calculating paper similarity
303 Once the decisions have been extracted and validated, this opens up a structured data for analyzing these information.
304 For example, we can compare whether author’s choices at different times changes, or across decisions varies at different
305 regions. In this section, we present a method to calculate paper similarity based on the decisions shared in the paper
306 pairs. The goal is to construct a distance metric based on similarity of the decision choice among papers that could
307 be further used for clustering paper based on choices made by different authors in the literature. An overview of the
308 method is illustrated in Figure 2.
309
310
311
312 Manuscript submitted to ACM

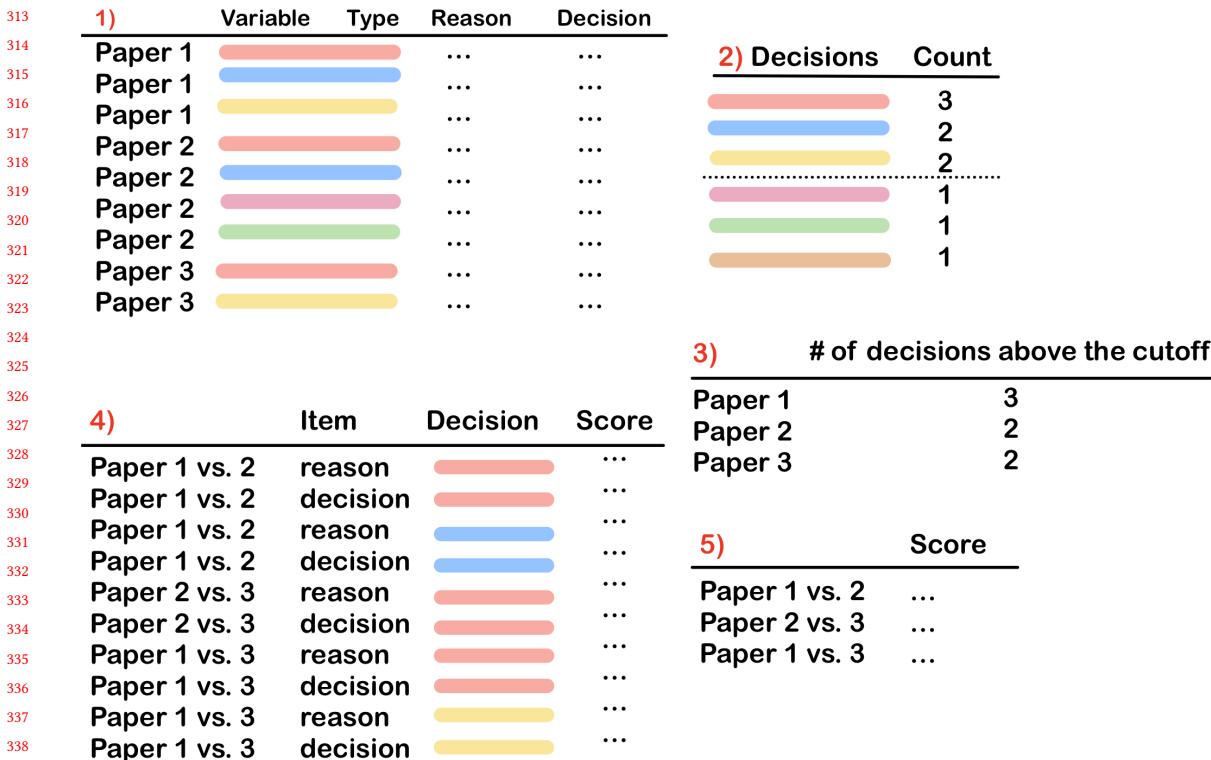


Fig. 2. Workflow for calculating paper similarity based on decision choices: (1) standardize variable names, (2) identify most frequent variable-type decisions across all papers, (3) identify papers with at least x identified decisions, (4) calculate decisions similarity score on the *decision* and *reason* 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”

- 365 • **time**: “date”, “time”, “trends”, “trend”

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

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

391 5 Results

392 5.1 Air pollution mortality modelling

- 393 • Given examples of the failure of LLM models for parsing and examples where authors are unclear about the
 394 delivery

395 The results follows examines [x] papers for modelling the effect of particulate matters on mortality based on Gemini
 396 for parsing the decision choices. The results from Anthropic Claude is reported in Section 5.2.

397 Specify how much of validation and review has been done

401 Decision quality summary

- 402 • missingness of the reason and decisions for the paper - how often papers report decisions
- 403 • look at for one type of decision (time) - what are the choices made by different papers
- 404 • look at whether decisions changes across time (cluster diagram with year)
- 405 • Visualize the decision database: apply clustering algorithm and visualize the clusters
- 406 • a characterization of the field, what are the common variables included, what smoothing methods are used,
 407 what are the options for temporal lags often considered, how are models generally estimated spatially.
- 408 • For lee2006association, it is not clear what specific smoothing method the sentence “smooth function of the
 409 day of study” refers to.

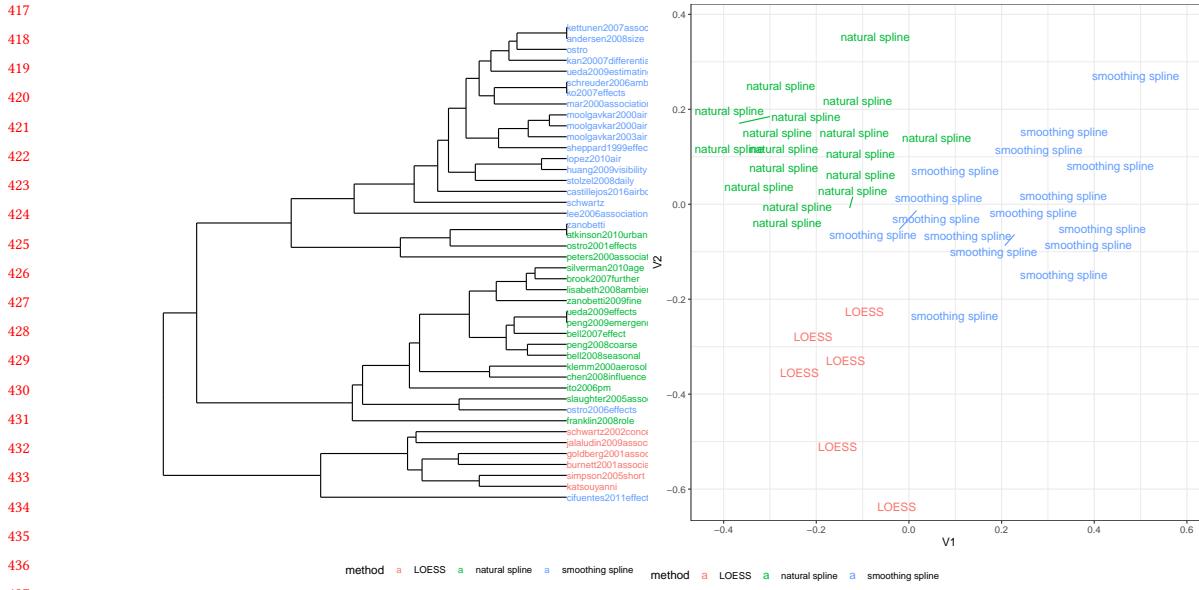


Fig. 3. bla bla bla

5.2 Sensitivity analysis

In this section, we examine the reproducibility for using LLMs for text extraction tasks in Section 5.2.1, discrepancies between different LLM models: Gemini (`gemini-2.0-flash`) and Claude (`claude-3-7-sonnet-latest`) in Section 5.2.2, and the sensitivity of our paper similarity calculation pipeline to the choice of text model used for computing decision similarity scores in Section 5.2.3.

5.2.1 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 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 2 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 2. 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

469 Table 3 summarizes the number of differences observed in each pairwise comparison. Among all comparisons, 80%
 470 produce the identical text in reason and decision. The discrepancies come from the following reasons:
 471

- 472 • Gemini extracted different length for the same decision, e.g. in Kan et al. [25], some runs may extract “singleday
 473 lag models underestimate the cumulative effect of pollutants on mortality 2day moving average **of current**
 474 **and previous day concentrations** (lag=01)”, while others extract “singleday lag models underestimate the
 475 cumulative effect of pollutants on mortality 2day moving average (lag=01)”. Similarity, for decisions, some runs
 476 may yield “10 df for total mortality”, while other runs yield “10 df”. Similar extraction appears in Breitner et al.
 477 [9].
- 478 • Gemini fails to extract reasons in some runs but not others, e.g. in Burnett et al. [10], the first run generates
 479 NAs in the reasons, but the remaining four runs are identical. In Ueda et al. [39] and Castillejos et al. [12], runs
 480 1 and 5 fail to extract the reasons and produce the same incomplete version, whereas runs 2, 3, and 4 produce
 481 accurate versions with reasons populated.

482 Table 3. Number of differences in the reason and decision fields across Gemini runs for papers with consistent number of decisions
 483 across runs.

484 Num. of difference	485 Count	486 Proportion (%)
487 0	488 358	489 79.73
490 1	491 12	492 2.67
493 2	494 8	495 1.78
496 3	497 0	498 0.00
499 4	500 24	501 5.35
502 5	503 12	504 2.67
505 6	506 3	507 0.67
508 7	509 0	510 0.00
511 8	512 10	513 2.23
514 9	515 6	516 1.34
517 10	518 10	519 2.23
520 11	521 6	522 1.34
523 Total		524 100.00

525 5.2.2 *LLM models*. Reading text from PDF document requires Optical Character Recognition (OCR) to convert images
 526 into machine-readable text, which currently is only supported by Anthropic Claude (claude-3-7-sonnet-latest) and
 527 Google Gemini (gemini-2.0-flash).

528 We compare the number of decisions extracted by Claude and Gemini across all 62 papers in Figure 4. Each point
 529 represents a paper, with the x- and y-axes showing the number of decisions extracted by Claude and Gemini, respectively.
 530 The dashed 1:1 line marks where both models extract the same number of decisions. Most points fall below this line,
 531 indicating that Claude extracts more decisions – often from data pre-processing or secondary data analysis steps
 532 requiring more manual validation – whereas Gemini focuses more on modelling choices relevant to our analysis. Some
 533 of these decisions captured by Claude are

- 534 • the definition of “cold day” and “hot day” indicators in Dockery et al. [16] (“defined at the 5th/ 95th percentile”),

- 521 • the choice to summarize NO₂, O₃, and SO₂ using a “24 hr average on variable” in Huang et al. [21], and
522 • the definition of black smoke and in Katsouyanni et al. [26] for secondary analysis (“restrict to days with BS
523 concentrations below 150 $\mu\text{g}/\text{m}^2$ ”).

524 Gemini sometimes also include irrelevant decisions, such as in Mar et al. [31], where secondary analysis choices like
525 “0-4 lag days” for air pollution exposure variables (CO, EC, K_S, NO₂, O₃, OC, Pb, S, SO₂, TC, Zn) are captured. However,
526 these cases are less frequent, resulting in outputs with less noise overall.

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

532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572

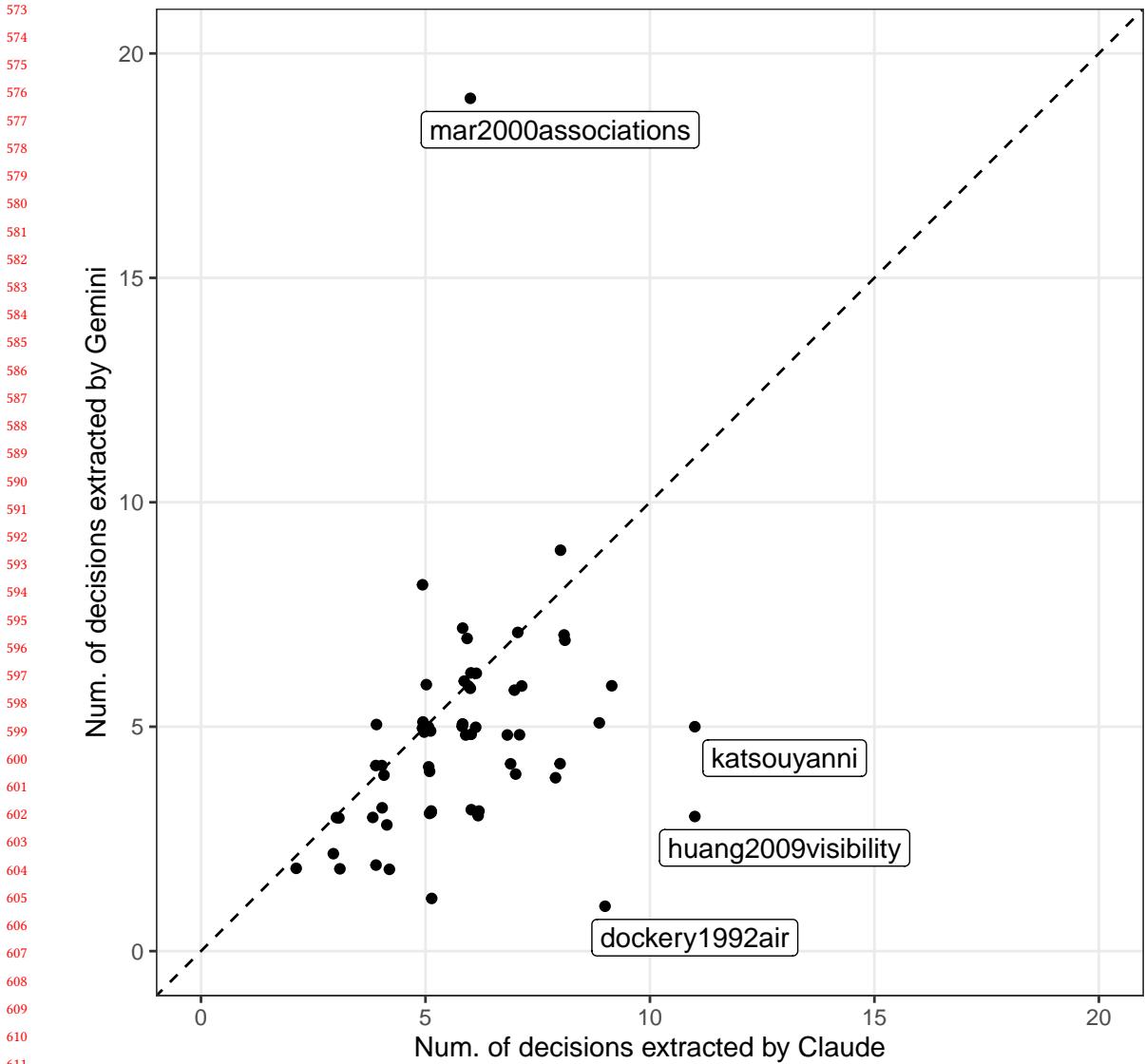


Fig. 4. Comparison of decisions extracted by Claude and Gemini. Each point represents a paper, with the x- and y-axes showing the number of decisions extracted by Claude and Gemini, respectively. The dashed 1:1 line marks where both models extract the same number of decisions. Most points fall below this line, suggesting Claude extracts more decisions – often including noise from data pre-processing or secondary data analysis steps – which requires additional manual validation.

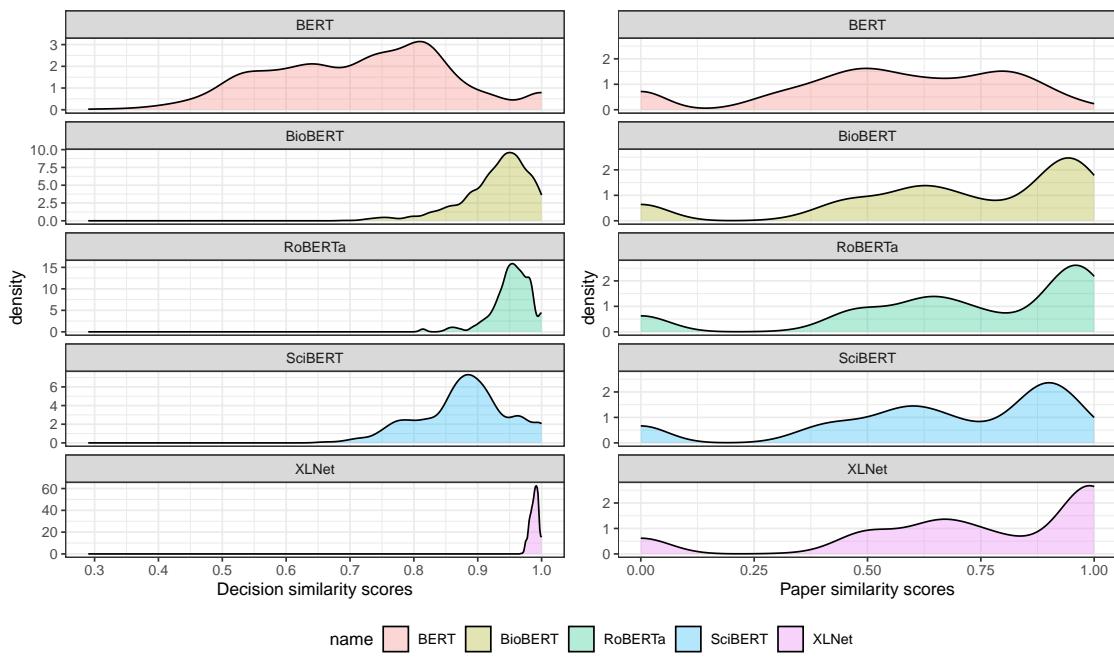
5.2.3 *Text model*. We have conducted sensitivity analysis on the text model for obtaining the decision similarity score from the Gemini outputs. The tested language models tested include

- 621 1) BERT by Google [15],
- 622 2) RoBERTa by Facebook AI [30], trained on a larger dataset (160GB v.s. BERT's 15GB),
- 623 3) XLNnet by Google Brain [42], and

625 two domain-trained BERT models:

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

629 Figure 5 presents the distribution of the decision similarity (left) and paper similarity (right) for each text model.
 630 At decision level, the BERT model produces the widest variation across all five models, while the similarity scores
 631 from XLNet are all close to 1. These scores are not comparable across models since the difference of the underlying
 632 transformer architecture. However, the paper similarity scores from each model are comparable and Figure 6 shows the
 633 multi-dimensional scaling (MDS) of the paper similarity scores from each text model: all showing a similar clustering
 634 pattern of the three main smoothing methods.
 635



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

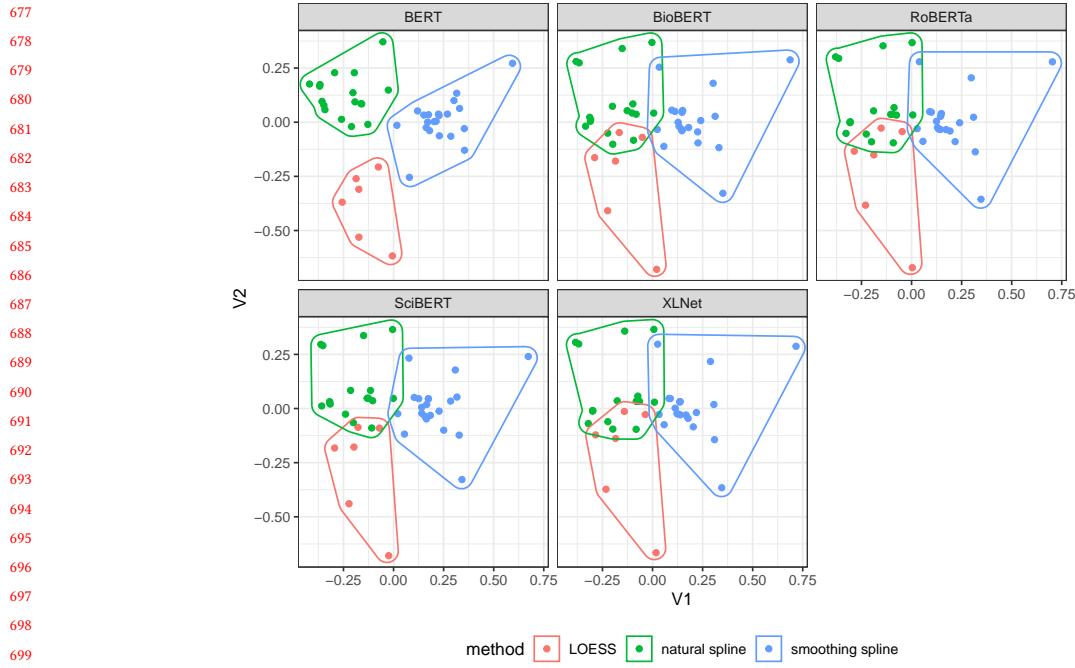


Fig. 6. The multi-dimensional scaling (MDS) of the paper similarity scores from each text model: all showing a similar clustering pattern of the three main smoothing methods. The points are colored by the most common method used in the paper, and the hulls are drawn around each method group.

6 Discussion

- Address how sensitivity analysis is/ is not relevant
- 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.
- Validation of the output:

the nature of the task: Our task involve a reasoning component in that it requires causal reasoning to identify the decisions made by the authors, and its justification/ rationale, rather than purely summarizing the text through pattern-matching.

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.bmjjournals.org/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 Barilaro, 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 Gleerean, 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 Hogeweegen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacobolla, 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 Nilsson, 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-Porras, 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>.

- [781] [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>.
- [782] [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>.
- [783] [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/>.
- [784] [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>.
- [785] [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>.
- [786] [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.
- [787] [19] Elliot Gould, Hannah S. Fraser, Timothy H. Parker, Shinichi Nakagawa, Simon C. Griffith, Peter A. Veski, 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 Blaard, 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, Xuechen Ge, William L. Geary, Charly Géron, Marc Gilles, Antje Girndt, Daniel Glikzman, 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-Chautre, Tanner J. Howard, Bin-Yan Hsu, Fabiola Iannarilli, Esperanza C. Iranzo, Erik N. K. Iverson, Saheed Olade Jimoh, Douglas H. Johnson, Martin Johnsson, Jesse Jorna, Tommaso Jucker, Martin Jung, Ineta Kačergyté, Oliver Kaltz, Alison Ke, Clint D. Kelly, Katharine Keegan, Friedrich Wolfgang Keppler, Alexander K. Killion, Dongmin Kim, David P. Kochan, Peter Korsten, Shan Kothari, Jonas Kuppler, Jillian M. Kusch, Małgorzata 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. Pruitt, Veronarindra Ramananjato, Finaritra Tolotra Randimbiharison, 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 Ruuskamen, 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. Schröder, 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 Slim, 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 Eckhoven, 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, Paweł 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.

- 833 *BMC Biology*, 23(1):35, 02 2025. doi: 10.1186/s12915-024-02101-x. URL <https://doi.org/10.1186/s12915-024-02101-x>.
- 834 [20] Florian Heimerl, Qi Han, Steffen Koch, and Thomas Ertl. Citerivers: Visual analytics of citation patterns. *IEEE Transactions on Visualization and*
 835 *Computer Graphics*, 22(1):190–199, 01 2016. doi: 10.1109/TVCG.2015.2467621. URL <https://ieeexplore.ieee.org/document/7192685/authors>.
- 836 [21] Wei Huang, Jianguo Tan, Haidong Kan, Ni Zhao, Weimin Song, Guixiang Song, Guohai Chen, Lili Jiang, Cheng Jiang, Renjie Chen, and Bingheng
 837 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.
 838 scitotenv.2009.02.019. URL <https://linkinghub.elsevier.com/retrieve/pii/S004896970900165X>.
- 839 [22] Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin
 840 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:
 841 <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecin.12992>.
- 842 [23] Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Jian Chen, and Torsten Möller. Visualization as seen through its research paper keywords. *IEEE*
 843 *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>.
- 844 [24] Alex Kale, Matthew Kay, and Jessica Hullman. Decision-making under uncertainty in research synthesis: Designing for the garden of forking
 845 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>.
- 846 [25] Haidong Kan, Stephanie J. London, Guohai Chen, Yunhui Zhang, Guixiang Song, Naiqing Zhao, Lili Jiang, and Bingheng Chen. Differentiating the
 847 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.
 848 2006.12.001. URL <https://www.sciencedirect.com/science/article/pii/S0160412006002108>.
- 849 [26] Klea Katsouyanni, Giota Touloumi, Evangelia Samoli, Alexandros Gryparis, Alain Le Tertre, Yannis Monopolis, Giuseppe Rossi, Denis Zmirou,
 850 Ferran Ballester, Azedine Boumghar, Hugh Ross Anderson, Bogdan Wojtyniak, Anna Paldy, Rony Braunstein, Juha Pekkanen, Christian Schindler,
 851 and Joel Schwartz. Confounding and effect modification in the short-term effects of ambient particles on total mortality: Results from 29 european
 852 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.
- 853 [27] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical
 854 language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 02 2020. doi: 10.1093/bioinformatics/btz682. URL
 855 <https://academic.oup.com/bioinformatics/article/36/4/1234/5566506>.
- 856 [28] Jiali Liu, Nadia Boukhelifa, and James R. Eagan. Understanding the Role of Alternatives in Data Analysis Practices. *IEEE Transactions on Visualization*
 857 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>.
- 858 [29] Yang Liu, Tim Althoff, and Jeffrey Heer. Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end
 859 data analysis. *CHI ’20*, page 1–14, New York, NY, USA, 04 2020. Association for Computing Machinery. doi: 10.1145/3313831.3376533. URL
 860 <https://dl.acm.org/doi/10.1145/3313831.3376533>.
- 861 [30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.
 862 Roberta: A robustly optimized bert pretraining approach. doi: 10.48550/arXiv.1907.11692.
- 863 [31] T F Mar, G A Norris, J Q Koenig, and T V Larson. Associations between air pollution and mortality in phoenix, 1995–1997. *Environmental*
 864 *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:
 865 Environmental Health Perspectives.
- 866 [32] Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. Vitality: Promoting serendipitous discovery of academic literature with
 867 transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):486–496, 01 2022. doi: 10.1109/TVCG.2021.3114820.
 868 URL <https://ieeexplore.ieee.org/document/9552447/>.
- 869 [33] Bart Ostro, Rachel Broadwin, Shelley Green, Wen-Ying Feng, and Michael Lipsett. Fine particulate air pollution and mortality in nine california
 870 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.
- 871 [34] Abhraneel Sarma, Alex Kale, Michael Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. multiverse: Multiplexing
 872 alternative data analyses in r notebooks (version 0.6.2). *OSF Preprints*, 2021. URL <https://github.com/MUCollective/multiverse>.
- 873 [35] Marko Sarstedt, Susanne J. Adler, Christian M. Ringle, Gyeongcheol Cho, Adamantios Diamantopoulos, Heungsun Hwang, and Benjamin D.
 874 Liengaard. Same model, same data, but different outcomes: Evaluating the impact of method choices in structural equation modeling. *Journal of*
 875 *Product Innovation Management*, 41(6):1100–1117, 2024. doi: 10.1111/jipim.12738. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/jipim.12738>.
 876 _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jipim.12738>.
- 877 [36] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtry, Š. Bahník, F. Bai, C. Barnard, E. Bonnier, R. Carlsson, F. Cheung,
 878 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,
 879 T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Höglén, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei,
 880 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,
 881 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,
 882 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
 883 884 Manuscript submitted to ACM

- 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.
- [37] 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>.
- [38] 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>.
- [39] 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>.
- [40] 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>.
- [41] 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.
- [42] 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.
- [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. Futzng and moseyng: 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. Ketz, 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.bmjjournals.org/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 Barilaro, 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 Durneze, 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, Gabriele Herman, Stephan Heinrich, Felix Hoffstaedter, Jeremy Hogeveen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacobella, Alexandru D. Iordan, Peder M. Isager, Ayse I. Isik, Andrew Jahn, Matthew R. Johnson, Tom Johnstone, Michael J. E. Joseph, Anthony C. Julian, 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 Lauharatananahirun, 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.

- 937 Samanez-Larkin, Emilio Sanz-Morales, Margaret L. Schlichting, Douglas H. Schultz, Qiang Shen, Margaret A. Sheridan, Jennifer A. Silvers, Kenny
938 Skagerlund, Alec Smith, David V. Smith, Peter Sokol-Hessner, Simon R. Steinkamp, Sarah M. Tashjian, Bertrand Thirion, John N. Thorp, Gustav
939 Tinghög, Loreen Tisdall, Steven H. Tompson, Claudio Toro-Serey, Juan Jesus Torre Tresols, Leonardo Tozzi, Vuong Truong, Luca Turella, Anna E.
940 van 't Veer, Tom Verguts, Jean M. Vettel, Sagana Vijayarajah, Khoi Vo, Matthew B. Wall, Wouter D. Weeda, Susanne Weis, David J. White, David
941 Wisniewski, Alba Xifra-Porras, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S. L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E.
942 Nichols, Russell A. Poldrack, and Tom Schonberg. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):
943 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
944 Wichmann, and Annette Peters. Short-term mortality rates during a decade of improved air quality in erfurt, germany. *Environmental Health
945 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
946 Health Perspectives.
- [10] Richard T. Burnett, Sabit Cakmak, Mark E. Raizenne, David Stieb, Renaud Vincent, Daniel Krewski, Jeffrey R. Brook, Owen Philips, and Haluk
947 Ozkaynak. The association between ambient carbon monoxide levels and daily mortality in toronto, canada. *Journal of the Air & Waste Management
948 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. Asso-
951 ciations between short-term changes in nitrogen dioxide and mortality in canadian cities. *Archives of Environmental Health: An International
952 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:
953 <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.
955 *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>.
956 Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/0895-8378.1987.11463182>.
- [13] Chaomei Chen. Citospace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society
957 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>.
958 _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.
960 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
963 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.
965 *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
967 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
971 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.
972 Abbey-Lee, Jessica K. Abbott, Luis A. Aguirre, Carles Alcaraz, Irith Aloni, Drew Altschul, Kunal Arekar, Jeff W. Atkins, Joe Atkinson, Christopher M.
973 Baker, Meghan Barrett, Kristian Bell, Suleiman Kehinde Bello, Iván Beltrán, Bernd J. Berauer, Michael Grant Bertram, Peter D. Billman, Charlie K.
974 Blake, Shannon Blake, Louis Bliard, Andrea Bonisioli-Alquati, Timothée Bonnet, Camille Nina Marion Bordes, Aneesh P. H. Bose, Thomas Botterill-
975 James, Melissa Anna Boyd, Sarah A. Boyle, Tom Bradfer-Lawrence, Jennifer Bradham, Jack A. Brand, Martin I. Brengdahl, Martin Bulla, Luc Bussière,
976 Ettore Camerlenghi, Sara E. Campbell, Leonardo L. F. Campos, Anthony Caravaggi, Pedro Cardoso, Charles J. W. Carroll, Therese A. Catanach,
977 Xuan Chen, Heung Ying Janet Chik, Emily Sarah Choy, Alec Philip Christie, Angela Chuang, Amanda J. Chunco, Bethany L. Clark, Andrea Contina,
978 Garth A. Covernton, Murray P. Cox, Kimberly A. Cressman, Marco Crotti, Connor Davidson Crouch, Pietro B. D'Amelio, Alexandra Allison
979 de Sousa, Timm Fabian Döbert, Ralph Dobler, Adam J. Dobson, Tim S. Doherty, Szymon Marian Drobniaik, Alexandra Grace Duffy, Alison B. Duncan,
980 Robert P. Dunn, Jamie Dunning, Trishna Dutta, Luke Eberhart-Hertel, Jared Alan Elmore, Mahmoud Medhat Elsherif, Holly M. English, David C.
981 Ensminger, Ulrich Rainer Ernst, Stephen M. Ferguson, Esteban Fernandez-Juricic, Thalita Ferreira-Arruda, John Fieberg, Elizabeth A. Finch, Evan A.
982 Fiorenza, David N. Fisher, Amélie Fontaine, Wolfgang Forstmeier, Yoan Fourcade, Graham S. Frank, Cathryn A. Freund, Eduardo Fuentes-Lillo,
983 Sara L. Gandy, Dustin G. Gannon, Ana I. García-Cervigón, Alexis C. Garretson, Xuezhen Ge, William L. Geary, Charly Géron, Marc Gilles, Antje
984 Girndt, Daniel Gliksman, Harrison B. Goldspiel, Dylan G. E. Gomes, Megan Kate Good, Sarah C. Goslee, J. Stephen Gosnell, Eliza M. Grames, Paolo
985 Gratton, Nicholas M. Grebe, Maaike Griffioen, Daniel M. Griffith, Frances J. Griffith, Jake J. Grossman, Ali Güncan, Stef Haesen,
986 James G. Hagan, Heather A. Hager, Jonathan Philo Harris, Natasha Dean Harrison, Sarah Syedia Hasnain, Justin Chas Havird, Andrew J. Heaton,
987 María Laura Herrera-Chaustrre, Tanner J. Howard, Bin-Yan Hsu, Fabiola Iannarilli, Esperanza C. Iranzo, Erik N. K. Iverson, Saheed Olaide Jimoh,
988 Douglas H. Johnson, Martin Johnsson, Jesse Jorna, Tommaso Jucker, Martin Jung, Ineta Kačergytė, Oliver Kaltz, Alison Ke, Clint D. Kelly, Katharine

- 989 Keegan, Friedrich Wolfgang Keppeler, Alexander K. Killion, Dongmin Kim, David P. Kochan, Peter Korsten, Shan Kothari, Jonas Kuppler, Jillian M.
 990 Kusch, Małgorzata Lagisz, Kristen Marianne Lalla, Daniel J. Larkin, Courtney L. Larson, Katherine S. Lauck, M. Elise Lauterbur, Alan Law, Don-Jean
 991 Léandri-Breton, Jonas J. Lembrechts, Kiara L'Herpiniere, Eva J. P. Lievens, Daniela Oliveira de Lima, Shane Lindsay, Martin Luquet, Ross MacLeod,
 992 Kirsty H. Macphie, Kit Magellan, Magdalena M. Mair, Lisa E. Malm, Stefano Mammola, Caitlin P. Mandeville, Michael Manhart, Laura Milena
 993 Manrique-Garzon, Elina Mäntylä, Philippe Marchand, Benjamin Michael Marshall, Charles A. Martin, Dominic Andreas Martin, Jake Mitchell
 994 Martin, April Robin Martinig, Erin S. McCallum, Mark McCauley, Sabrina M. McNew, Scott J. Meiners, Thomas Merkling, Marcus Michelangeli,
 995 Maria Moiron, Bruno Moreira, Jennifer Mortensen, Benjamin Mos, Taofeek Olatunbosun Muraina, Penelope Wrenn Murphy, Luca Nelli, Petri
 996 Niemelä, Josh Nightingale, Gustav Nilsonne, Sergio Nolazco, Sabine S. Nooten, Jessie Lanterman Novotny, Agnes Birgitta Olin, Chris L. Organ,
 997 Kate L. Ostevik, Facundo Xavier Palacio, Matthieu Paquet, Darren James Parker, David J. Pascall, Valerie J. Pasquarella, John Harold Paterson, Ana
 998 Payo-Payo, Karen Marie Pedersen, Grégoire Perez, Kayla I. Perry, Patrice Pottier, Michael J. Proulx, Raphaël Proulx, Jessica L. Pruitt, Veronarindra
 999 Ramananjato, Finaritra Tolotra Randimbirison, Onja H. Razafindratisima, Diana J. Rennison, Federico Riva, Sepand Riyahi, Michael James Roast,
 1000 Felipe Pereira Rocha, Dominique G. Roche, Cristian Román-Palacios, Michael S. Rosenberg, Jessica Ross, Freya E. Rowland, Deusdedith Rugemalila,
 1001 Avery L. Russell, Suvi Ruuskamen, Patrick Saccone, Asaf Sadeh, Stephen M. Salazar, Kris Sales, Pablo Salmón, Alfredo Sánchez-Tójar, Leticia Pereira
 1002 Santos, Francesca Santostefano, Hayden T. Schilling, Marcus Schmidt, Tim Schmoll, Adam C. Schneider, Allie E. Schrock, Julia Schroeder, Nicolas
 1003 Schtickzelle, Nick L. Schultz, Drew A. Scott, Michael Peter Scroggie, Julie Teresa Shapiro, Nitika Sharma, Caroline L. Shearer, Diego Simón, Michael I.
 1004 Sitvarin, Fabrício Luiz Skupien, Heather Lea Slimm, Grania Polly Smith, Jeremy A. Smith, Rahel Sollmann, Kaitlin Stack Whitney, Shannon Michael
 1005 Still, Erica F. Stuber, Guy F. Sutton, Ben Swallow, Conor Claverie Taff, Elina Takola, Andrew J. Tanentzap, Rocío Tarjuelo, Richard J. Telford,
 1006 Christopher J. Thawley, Hugo Thierry, Jacqueline Thomson, Svenja Tidau, Emily M. Tompkins, Claire Marie Tortorelli, Andrew Trlica, Biz R.
 1007 Turnell, Lara Urban, Stijn Van de Vondel, Jessica Eva Megan van der Wal, Jens Van Eckhoven, Francis van Oordt, K. Michelle Vanderwel, Mark C.
 1008 Vanderwel, Karen J. Vanderwolf, Juliana Vélez, Diana Carolina Vergara-Florez, Brian C. Verrelli, Marcus Vinícius Vieira, Nora Villamil, Valerio
 1009 Vitali, Julien Vollering, Jeffrey Walker, Xanthe J. Walker, Jonathan A. Walter, Paweł Waryszak, Ryan J. Weaver, Ronja E. M. Wedegärtner, Daniel L.
 1010 Weller, and Shannon Whelan. Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology.
 1011 *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 apheas2 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] 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/>.
- [29] 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>.
- 1040 Manuscript submitted to ACM

- 1041 [30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.
 1042 Roberta: A robustly optimized bert pretraining approach. doi: 10.48550/arXiv.1907.11692.
- 1043 [31] T F Mar, G A Norris, J Q Koenig, and T V Larson. Associations between air pollution and mortality in phoenix, 1995-1997. *Environmental
 1044 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:
 1045 Environmental Health Perspectives.
- 1046 [32] Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. Vitality: Promoting serendipitous discovery of academic literature with
 1047 transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):486–496, 01 2022. doi: 10.1109/TVCG.2021.3114820.
 1048 URL <https://ieeexplore.ieee.org/document/9552447/>.
- 1049 [33] Bart Ostro, Rachel Broadwin, Shelley Green, Wen-Ying Feng, and Michael Lipsett. Fine particulate air pollution and mortality in nine california
 1050 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.
- 1051 [34] Abhraneel Sarma, Alex Kale, Michael Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. multiverse: Multiplexing
 1052 alternative data analyses in r notebooks (version 0.6.2). *OSF Preprints*, 2021. URL <https://github.com/MUCollective/multiverse>.
- 1053 [35] Marko Sarstedt, Susanne J. Adler, Christian M. Ringle, Gyeongcheol Cho, Adamantios Diamantopoulos, Heungsun Hwang, and Benjamin D.
 1054 Lienggaard. Same model, same data, but different outcomes: Evaluating the impact of method choices in structural equation modeling. *Journal of
 1055 Product Innovation Management*, 41(6):1100–1117, 2024. doi: 10.1111/jpim.12738. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/jpim.12738>.
 1056 _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jpim.12738>.
- 1057 [36] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtry, Š. Bahnik, F. Bai, C. Barnard, E. Bonnier, R. Carlsson, F. Cheung,
 1058 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,
 1059 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,
 1060 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,
 1061 E. Robusto, H. Roderique, A. Sandberg, E. Schläuter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford,
 1062 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
 1063 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.
- 1064 [37] Jan Simson, Fiona Draxler, Samuel Mehr, and Christoph Kern. Preventing harmful data practices by using participatory input to navigate the
 1065 machine learning multiverse. CHI '25, page 1–30, New York, NY, USA, 04 2025. Association for Computing Machinery. doi: 10.1145/3706598.3713482.
 1066 URL <https://dl.acm.org/doi/10.1145/3706598.3713482>.
- 1067 [38] Imad Tbahriti, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. Using argumentation to retrieve articles with similar citations: An
 1068 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>.
- 1069 [39] Kayo Ueda, Nitta ,Hiroshi ,Ono ,Masaji , , and Ayano Takeuchi. Estimating mortality effects of fine particulate matter in japan: A comparison of time-
 1070 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.
 1071 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>.
- 1072 [40] 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>.
- 1073 [41] Hadley Wickham, Joe Cheng, and Aaron Jacobs. *ellmer: Chat with Large Language Models*, 2025. URL <https://CRAN.R-project.org/package=ellmer>.
 1074 R package version 0.1.1.
- 1075 [42] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for
 1076 language understanding. doi: 10.48550/arXiv.1906.08237.
- 1077
 1078
 1079
 1080
 1081
 1082
 1083
 1084
 1085
 1086
 1087
 1088
 1089
 1090
 1091
 1092