Numbers in the News: A Survey

Robert L. Raymond[1] and Milo Schield[2]
[1] Emeritus, University of Saint Thomas, St. Paul, MN
[2] Augsburg College, Minneapolis, MN

Abstract

This report presents the prevalence of statistical terms and ideas in news stories. This involves (1) a judgment analysis of 160 articles and (2) a computer-match analysis of 899 articles. The judgment analysis of 160 articles involved 73 traits. These included (a) elements of study design, (b) statistical measures, and (c) types of grammar used to describe and compare ratios. Of these articles, 27% involve assembly in constructing categories or measures and 62% present associations that imply causation. Using judgment we find the following influences plausible: 42% influenced by confounding, 17% by assembly, 11% by bias and 9% by random effects. About 79% of the articles are based on samples. Of those articles, 89% gave the sample size but only 3% stated that the sample was random and none gave a p-value. Many studies described populations, for example all patients treated for a condition at predetermined hospitals during a fixed time period. Such articles never discussed why these results might or might not generalize to a larger population. The computer-match of 899 number-based news articles calculated the prevalence of 231 statistical terms. In this set, 19% use the word “significant,” but only 3% use “statistically significant” and none use “statistical significance”. Of these articles, over 90% used comparisons, 74% used percent grammar, 67% used causal grammar, 61% used association grammar, 55% used chance grammar, 32% used rate grammar, 13% used percentage grammar and 10% used confounding grammar. This empirical data should be useful in determining topics and their emphasis in teaching statistical literacy as applied critical thinking.

1. Introduction

An earlier article, Schield & Schield (2007), connected the use of statistics from the media to the goal of statistical literacy and the recommendations in the GAISE College Report.

Statistics instructors should model the use of real data when they design their courses. If statistical literacy is to be empirically based, the use of numbers in everyday venues must be analyzed. The goal of this paper is to survey the use of various kinds of statistics and associated factors in news articles.

1.2 Articles Selected and Analyzed

The articles studied appeared primarily on the front page or health sections of Yahoo. Most were released between 8/2005 and 4/2008. The news articles are typically one or two pages. Articles involving sports, weather, stock prices or original research studies were excluded. This source does not include tables or graphs in its articles, so we have not dealt with any issues related to them.

News articles were considered if they involve numbers and they: have “study,” “survey” or “report” in the title, involve or reference a study, survey or report, involve diagnostic tests (medical or otherwise), involve longitudinal data or subject manipulation, involve random assignment or random selection, involve a sample, sample size or margin of error, have “significantly” or “(in)significant” in the text, involve taking into account a confounder, or use statistics as evidence for causation.

1.3 Data Tabulation

Data was tabulated in two ways: subjectively and objectively. As will be seen, objective is not necessarily better.

Appendix A shows the data entry form for subjective analysis and indicates how each of the 73 fields was handled. Note that the choice and definitions of the statistical categories may result in some topics being omitted (non-exhaustive) while others may be obscured either by having two topics grouped under a single heading or by having one topic split between two headings (non-exclusive).

Each of the 899 articles in the database was objectively analyzed for the presence of 231 keywords. Unfortunately, this mechanical approach cannot distinguish between the expected and unexpected use of a keyword. For example, the word “associated” appears in 55% of these articles, but the phrase “associated press” appears in 30% of them. Thus the prevalence of keywords is an upper-limit on the prevalence in their expected use.

2. Overall Findings

Findings are presented in two groups. The first set is based on the subjective interpretation by the reviewer for 160 articles analyzed in 2008 as explained in Appendix A. The results of that analysis are shown in Appendix B and are summarized in Tables 1-7. The second set is based on mechanical matching of 231 terms with the content of the articles. Those results are shown in Appendix C and are discussed in Section 4.

Table 1: Prevalence of study design characteristics

79%	Mention using a sample
71%	Sample size
32%	Longitudinal
31%	Cohort based
26%	Controlled
13%	Subject manipulation
11%	Controlled by selection
9%	Additional factor controlled
4%	Subject blinded (placebo)
2%	Random assignment

Table 2: Prevalence of simpler numbers

71%	Percent (part-whole, % chg)
58%	Rates
30%	Numbers (counts or measures)
12%	Ratios (e.g., miles per gallon)
2%	Ranks or percentiles

Table 3: Prevalence of more complex numbers

4%	Slope
2%	Range

Table 4: Prevalence of arithmetic comparisons

65%	Quantitative compare
19%	Qualitative compare
2%	Cases attributed to
2%	%‘Attributed[able] to’

Table 5: Prevalence of ratio grammars.

34%	Percent grammar
31%	Rate grammar
23%	Chance grammar
1%	Percentage grammar

Table 6: Prevalence of sample-related characteristics

10%	“ unlikely due to chance ”
2%	“ random sample ”
1%	“ Margin of error ”

Table 7: Troublesome characteristics

42%	Confounding plausible
17%	Assembly plausible
11%	Bias plausible
9%	Random effects plausible

Schield (2000) analyzes the keywords and syntax used to describe and compare selected ratios. Percent grammar is identified by “percent of” or “percent is/are.” Rate grammar is identified by “per.” Rate includes prevalence (a part-whole ratio that doesn’t involve a time interval such as unemployment), incidence (a ratio per time interval such as “births per 10,000 women per year”) or a “velocity” (things per unit time such as “births per year”). Chance grammar involves chance, risk, odds, likelihood or probability. Percentage grammar involves “percentage of” or “percentage that/who”.

Statistical measures that never appeared in these news articles include “percentage explained by” and “p-value”.

3. Critical Thinking Characteristics

Tables 1 – 6 summarize various statistical characteristics. However they do not present two ways that statistics relate to critical thinking. One is how the statistics are constructed or assembled; the other is how these statistical associations are presented to imply causation. Table 7 tabulated the frequency of concerns related to these issues. Isaacson (2005) and Schield (2007) have argued that these concerns are pivotal in evaluating the nature and strength of any argument using the statistics. As will be shown, they are found in most news articles. Thus they should be central to any statistical literacy course.

Caution must be used in citing Appendix B because the numbers in some categories are small. It appears that assembly is most plausible for ordinal outcome measures, confounding most troubling with multinomial or binary outcomes, randomness for discrete, and bias if the outcome is multinomial or discrete.

3.1 Assembly or Construction of Statistics

Joel Best (2001) noted that all statistics are socially constructed in that they are selected, defined, counted/measured, summarized and presented by people that have interests in seeing a number be larger or smaller.

Of these articles, 27% use words that have considerable latitude in their definition or measure. We call such words instances of “assembly.” See Schield (2007). We judged that the results stated in 17% of the 160 articles may have been influenced by assembly. Here are examples of assembly from articles studied: scales or scores for medication schedule adherence, dysfunction, jealousy, loneliness, temperament, or movie profanity; grouping weights based on BMI; race; and types of church experiences. Other words lacking generally accepted definitions such as excessive, discrepancy, dangerous, rich, poor, unhealthy and dysfunctional appear.

3.2 Using Association to Imply Causation

Of the articles, 62% use words that imply or assert causation when it may be highly disputable. Implying causation can be done through action verbs (e.g., help, change, alter, increase, improve, save, prevent, reduce, cut, kill or hurt), adjectives: (e.g., harmful, safe or effective) or nouns (e.g., fighter, protection, defense). Examples of phrases implying causation taken from the articles studied include: "helps control," "would cut," "instilled," or "can save."

Asserting causation when it is highly disputable occurs often in popular-cause issues. For example, "Second-hand smoke *causes* cancer." In this instance randomized trials are unethical, before-after studies are not repeatable and – unlike actual smoking – the relative risks are so low that their susceptibility to confounding makes the causal claim highly disputable.

4. Text-Matching Results

Appendix C contains the results of machine-generated matching on 899 articles. These articles were selected on the same criteria as the 160 mentioned previously and included these 160 articles. Matching involved 231 words or phrases as listed in Appendix C where they are sorted by their prevalence. The 183 search words or phrases with non-zero prevalences were grouped into nine categories as follows:

1: Article type: Study (75.0%), Report (65.1%), Survey (15.9%), StudyTitle (15.4%), ReportTitle (1.4%), SurveyTitle (1.2%).

2: Cause/Association: Because (47.9%), Results (27.3%), Association (23.8%), Related (23.8%), Cause (21.9%), Associated (19.6%), Effect (19.0%), Effects (18.8%), Linked (16.1%), Causes (14.1%), Link (11.6%), Caused (10.2%), Factor (9.9%), Relationship (9.1%), Result (8.9%), EffectsOf (8.2%), CausedBy (6.1%), ResultsOf (4.8%), EffectOf (4.7%), Relation (3.2%), ResultOf (2.9%), ResultIn (2.7%), ResultsIn (2.2%), Resulted (1.4%), Relate (1.0%), ResultedIn (0.9%), Causal (0.8%), Correlated (0.8%), Correlate (0.4%).

3: Measures: Average (27.1%), Mean (7.1%), Range (6.3%), PercentagePoints (5.9%), RangeOf (3.6%), Ranked (2.3%), Median (1.6%), Ranks (0.9%), RangeFrom (0.8%), Percentile (0.6%), Rank (0.4%), Skewed (0.3%), Mode (0.2%), StdDev (0.2%), Percentiles (0.1%), Quartile (0.1%), PercentagePoint (0.1%), Outlier* (0.1%)

4: Ratios and Models: Percent (65.3%), Risk (47.9%), Likely (42.6%), PercentOf (42.3%), RiskOf (34.7%), LikelyTo (33.1%), Rate (25.4%), Per (16.7%), RateOf (12.3%), Percentage (12.0%), % (9.8%), %Of (7.9%), TheRate (7.9%), Chance (6.2%), Share (5.7%), PercentageOf (5.3%), TheRateOf (5.3%), Incidence (5.3%), Prevalence (5.2%), ThePercentage (4.7%), %Confidence (4.4%), IncidenceOf (4.1%), Attributed (3.9%), Percentages (3.9%), Odds (3.9%), Likelihood (3.9%), ThePercentageOf (3.8%), ChanceOf (3.8%), PrevalenceOf (3.7%), LikelihoodOf (3.1%), AttributedTo (3.0%), Ratio (2.9%), Probability (2.4%), ProbabilityOf (2.3%), OddsOf (1.8%), ShareOf (1.4%), ChanceTo (1.3%), RatioOf (1.3%), Attribute (1.2%), Fraction (1.1%), Attributable (1.0%), Attributes (1.0%), AttributableTo (1.0%), FractionOf (0.9%), PercentChance (0.8%), RiskThat (0.8%), Standardize* (0.4%), LikelihoodThat (0.4%), Regress* (0.3%), ChanceThat (0.3%), OddsThat (0.3%), RelativeRisk (0.2%), OddsRatio (0.2%), Normalize* (0.2%), ChanceFor (0.1%).

5: Comparatives: More (90.0%), MoreThan (46.8%), Less (41.7%), ErThan (32.1%), MoreLikely (24.9%), Times (22.8%), MoreLikelyTo (20.7%), LessThan (12.8%), LessLikely (10.7%), LessLikelyTo (9.2%), LikelyThan (8.5%), MoreLikelyThan (6.8%), TimesMore (6.6%), AsLikelyTo (4.1%), PercentMore (3.9%), PercentIncrease (2.6%), LessLikelyThan (2.4%), LikelyAs (2.3%), AsLikelyAs (2.0%), TimesAs (1.7%), NumMoreThan (1.3%), TimesLess (.8%), %More (.7%), PercentDecrease (.3%), TimesAsMuch (.3%), PercentMoreThan (.2%), PercentLessThan (.2%), TimesMoreThan (.2%), %Less (.1%), PercentLess (.1%), %Decrease (.1%), TimesLessThan (.1%).

6: Study design: Control (23.4%), RandomlyAssigned (2.1%), Experiment (1.8%), ControlGroup (1.4%), Randomized (1.1%), ControlOf (1.0%), ClinicalTrial (0.9%), ControlledStudy (0.4%), Longitudinal (0.4%), ObservationalStudy (0.2%), RandomlyAssign (0.2%).

7: Sampling and statistical inference: Population (19.7%), Significant (18.8%), Sample (8.0%), Error (6.6%), Sampling (5.8%), ConfidenceLevel (4.8%), 95%Confiden* (4.4%), RandomSample (2.7%), RandomlySampled

(2.7%), Representative (2.7%), StatisticallySignificant (2.7%), Sampled (0.4%), MarginOfError (0.4%), NotAsignificant (0.2%), DueToChance (0.1%), InSignificant (0.1%).

8: Bias: Placebo (3.2%), Bias (0.7%), PlaceboEffect (0.3%), Biased (0.2%).

9: Confounding: Account (8.5%), AccountFor (3.7%), IntoAccount (3.7%), AccountsFor (1.6%), ControlFor (0.8%), ControllingFor (0.7%), TakenIntoAccount (0.7%), TakeIntoAccount (0.4%), TakingIntoAccount (0.4%), AccountOf (0.3%), TakesIntoAccount (0.3%), ControlledFor (0.2%), Confounding (0.2%).

Note that comparatives, ratios and models are extremely common while bias and confounding are very uncommon. There are many comparatives that this search did not capture. E.g., prices/sales/profits up/down X%. Of those terms that can indicate ratios, percent was most common followed by risk, likely, rate, per, percentage, %, chance, percentage points, share, incidence, prevalence, odds, likelihood, ratio, probability, fraction, percentile, relative risk and odds ratio. Note that 19% of the articles used 'significant.' but only 3% used 'statistically significant.'

Since one cannot just add the individual percentages, some of these words and phrases were grouped together and their combined prevalence was determined:

- 67% Causal: "results", "cause", "causes", "caused", "causal", "result", "resulted", "effect", or "effects".
- 61% Associate: "association" or "associated*" which excludes "Associated Press"
- 74% Percent: "percent of/is/are", fraction or share.
- 55% Chance: "chance of/to/that", risk, odds, likelihood or probability.
- 32% Rate: "rate", "prevalence" or "incidence."
- 13% Percentage: "percentage of/who/that"
- 10% Confounding: "account", "take/ing into account", "account/ed/ing for" or "control/ed/ing for"

Care must be taken in interpreting these results. While a computer-match is accurate, the word may be used in a different context. Consider 'chance'; it can be used colloquially (he took a chance by quitting his job).

5. Evaluation

Comparison of subjective and computer analyses is difficult. The computer provides an upper bound. In some cases the numbers are quite close: subjectively, 26% of studies were considered controlled, while the total for all "control" words is about 28%. There is a large difference between the figures for percent, 34% to 65%. It might be due to nonstandard usage of "percent", or perhaps just differences between the set and the subset analyzed subjectively.

Note that few of the suggested extensions from Schield (2007) were adopted. Almost all of those suggestions involved a judgment. The goal of this paper was to reduce judgment – not increase it. Instead this paper did something not foreseen in Schield (2007). It did a machine-generated comparison of selected keywords with the texts of all articles involved. While this may overstate the prevalences, it serves as an objective upper-limit.

6. Conclusion

If statistical literacy is to be GAISE-based, it must focus on analyzing statistics in the news. Based on the prevalence of statistical terms and ideas found in this analysis of number-related news articles, topics in statistical literacy could be emphasized in the following order based on the rounded percentage of articles involved: 80% sampling and sample size, 70% grammar involving "rate," 70% percents and percent grammar, 65% comparisons made quantitatively, 30% longitudinal studies, 30% assembly, 30% controlled studies, 25% association-like words that imply causation, 25% "chance", 20% "significant," 10% experimental manipulation, 3% random assignment, 3% "statistically significant," 1% grammar using "percentage," 1% margin of error in surveys.

Just as Sherlock Holmes noted the "peculiar behavior" of the dog that did nothing in the night, our students should also be aware of information that is needed but missing. Knowing why a random sample is desirable and how to carry out a study so as to minimize bias and confounding is important perhaps most of all because this information is so often omitted. Assembly is never flagged as such by authors.

Note that a low prevalence in news articles should not be the sole criteria for determining to omit the associated statistical topic. Low-prevalence topics that may be important include phrases such as 'percentage' grammar (13%), 'statistically significant' (2.7%), 'margin of error' (0.4%), confounding (0.2%) and 'controlled for' (0.2%).

Acknowledgments

This paper is part of the W. M. Keck Statistical Literacy project: a project "to support the development of statistical literacy as an interdisciplinary curriculum in the liberal arts." The data[1] for this paper[2] is available.

References

Best, Joel (2001). "*Damned Lies and Statistics.*" University of California Press.[3]

Isaacson, Marc (2005). "Statistical Literacy: An Online Course at Capella University." JSM 2005 Section on Statistical Education, [CD-ROM] pp. 2244-2252.

Schield, Milo (2000). "Statistical Literacy: Difficulties in Describing and Comparing Rates and Percentages."[4] JSM 2000 Section on Statistical Education, American Statistical Association, pp. 76-81.

Schield, Milo (2007). "Teaching the Social Construction of Statistics."[5] Midwest Sociological Society, spring 2007.

Schield, Milo & Schield, Cynthia (2007). "Numbers in the News: A Survey." ASA 2007 Section on Statistical Education, [CD-ROM] pp. 2243-2250.[6]

APPENDIX A: DATA ENTRY

Figure 1 illustrates the data entry form that was manually entered for each article.

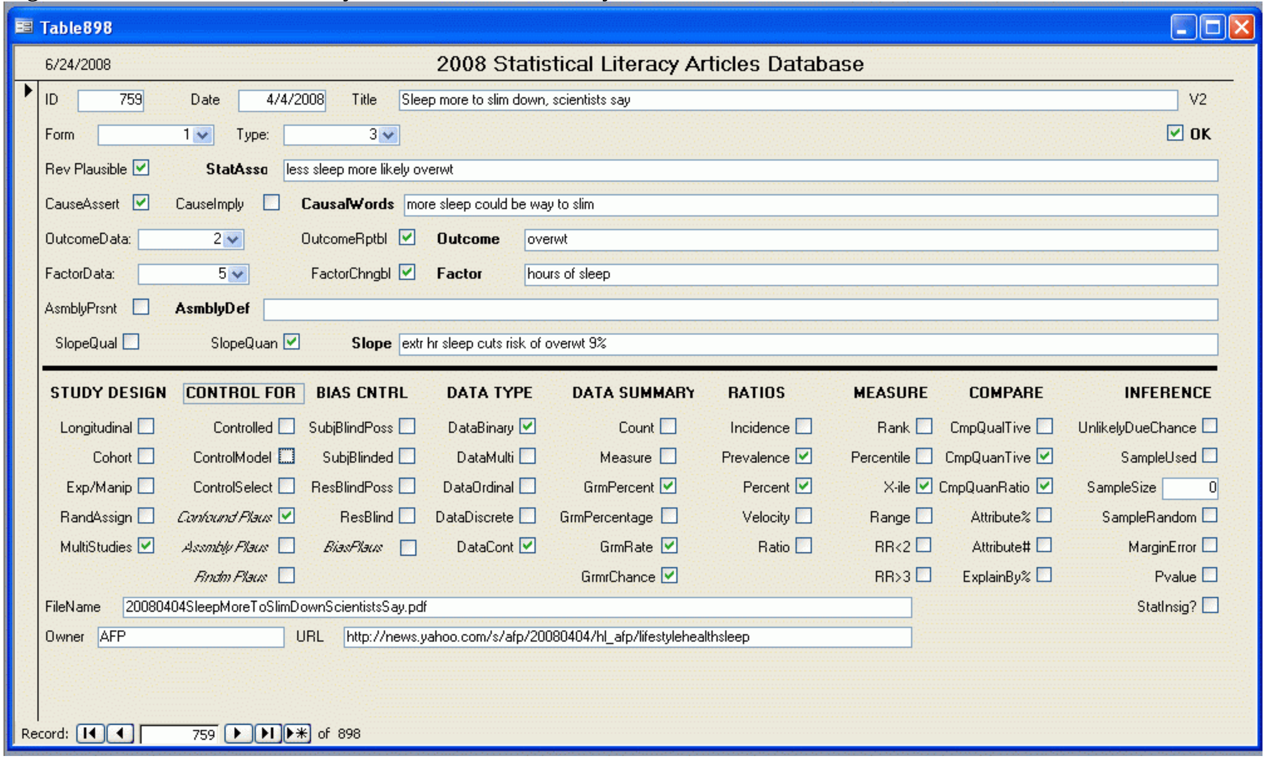

Table898

6/24/2008 2008 Statistical Literacy Articles Database

ID 759 Date 4/4/2008 Title Sleep more to slim down, scientists say V2

Form 1 Type: 3 ☑ OK

Rev Plausible ☑ StatAsso less sleep more likely overwt

CauseAssert ☑ CauseImply □ CausalWords more sleep could be way to slim

OutcomeData: 2 OutcomeRptbl ☑ Outcome overwt

FactorData: 5 FactorChngbl ☑ Factor hours of sleep

AsmblyPrsnt □ AsmblyDef

SlopeQual □ SlopeQuan ☑ Slope extr hr sleep cuts risk of overwt 9%

STUDY DESIGN	CONTROL FOR	BIAS CNTRL	DATA TYPE	DATA SUMMARY	RATIOS	MEASURE	COMPARE	INFERENCE
Longitudinal □	Controlled □	SubjBlindPoss □	DataBinary ☑	Count □	Incidence □	Rank □	CmpQualTive □	UnlikelyDueChance □
Cohort □	ControlModel □	SubjBlinded □	DataMulti □	Measure □	Prevalence ☑	Percentile □	CmpQuanTive ☑	SampleUsed □
Exp/Manip □	ControlSelect □	ResBlindPoss □	DataOrdinal □	GrmPercent ☑	Percent ☑	X-ile ☑	CmpQuanRatio ☑	SampleSize 0
RandAssign □	*Confound Flaus* ☑	ResBlind □	DataDiscrete □	GrmPercentage □	Velocity □	Range □	Attribute% □	SampleRandom □
MultiStudies ☑	*Assmbly Flaus* □	*BiasFlaus* □	DataCont ☑	GrmRate ☑	Ratio □	RR<2 □	Attribute# □	MarginError □
	Rndm Flaus □			GrmrChance ☑		RR>3 □	ExplainBy% □	Pvalue □
								StatInsig? □

FileName 20080404SleepMoreToSlimDownScientistsSay.pdf

Owner AFP URL http://news.yahoo.com/s/afp/20080404/hl_afp/lifestylehealthsleep

Record: 759 of 898

Figure 1: Data Entry Screen

[1] www.StatLit.org/xls/2008RaymondSchieldASA.xls
[2] www.StatLit.org/pdf/2008RaymondSchieldASA.pdf
[3] www.StatLit.org/Best.htm
[4] www.StatLit.org/pdf/2000SchieldASA.pdf
[5] www.StatLit.org/pdf/2007SchieldMSS.pdf
[6] www.StatLit.org/pdf/2007SchieldASA.pdf

NATURE AND INTERPRETATION OF FIELDS

These 75 fields (73 + ID and OK) were interpreted as follows:

- ID: Automatically generated by MS Access.
- Form: 1. News article; 2. Press release; 3. Detailed study
- Type: 1=Medical test, 2 = Survey, 3 = Study-related, 4 = Study, blank = Other. "Survey" refers to an article based on extracting interesting findings from a survey, not answering a predetermined question using a survey.

TEXTBOXES (Top of form)

- Date: Publication date shown in the article.
- Title: Taken as-is from the title of the article.
- Stat. association: one taken from the article.
- Causal words: Words that imply or state causation
- Outcome: Result of interest.
- Factor: Factor associated with outcome.
- AsmblyDef: Words w. plausible assembly in definition
- Slope: any regression-like relationship

NUMBER BOXES (Top of form)

- OutcomeData: 1=Multinomial, 2=Binary, 3=Ordinal, 4=Discrete, 5=Continuous
- FactorData: 1=Multinomial, 2=Binary, 3=Ordinal, 4=Discrete, 5=Continuous

CHECKBOXES: GENERAL (Top of form)

- OK: Article evaluation complete. Article OK.
- Reverse plausible: Difference in the results can cause the difference observed in the predictors.
- CauseAssert: Article asserts causation
- CauseImply: Article implies causation
- Outcome repeatable: Outcome event can be repeated for a given subject. (E.g., headache = Yes; Dying = No)
- FactorChangbl: Predictor can be readily changed for a given subject. (E.g., smoker = Yes; Male = No)
- AssmblyPresent: Assembly in presentation.
- SlopeQual: Slope given qualitatively. As X incr, Y incr.
- SlopeQuan: Rise/Run. E.g., Y incr by 2 if X incr by 1.

CHECKBOXES: STUDY DESIGN

- Longitudinal: Multiple measures of outcome over time. Not checked for a prospective or retrospective study that had only one measure of a continuing outcome, nor if the outcome was not repeatable for unit of analysis
- Cohort: Subjects are a closed group (some may drop out).
- Exp/Manip: Subjects intentionally manipulated.
- RndmAssign: Subjects randomly assigned.
- Multiple studies: Multiple studies referenced.

CHECKBOXES: CONTROL FOR

- Controlled: Article references multiple groups.
- ControlledModel: Control via model (regression)
- ControlledSelect: Control via selection
- *ConfoundPlaus?*: Confounding more likely than not.
- *AssmblyPlaus?*: Assembly more likely than not.
- *RndmPlaus*?: Influence of randomness likely

CHECKBOXES: BIAS CONTROL

- SubjBlindPoss?: Was it possible to blind the subject?
- SubjectBlind: Subjects blinded as to their group.
- ResBlindPoss?: Possible to blind the researcher?
- EvaluatorBlind: Evaluators blinded to a subject's group.
- BiasPlaus?: Could bias explain substantial part of result?

CHECKBOXES: DATA TYPE

- DataBinary: Includes some binary data
- DataMult: Includes some multinomial data
- DataOrdinal: Includes some ordinal data
- DataDiscrete: Includes some discrete data
- DataCont: Includes some continuous data

CHECKBOXES: DATA SUMMARY

- Count
- Measure
- GrmrPercent: part-whole measure
- GrmrPercentage: part-whole measure.
- GrmrRate: numerator/denominator measure
- GrmrChance:

CHECKBOXES: RATIOS

- Incidence: E.g., Deaths per 1,000 people per yr.
- Prevalence: E.g., Unemployment
- Percent: Uses 'per 100' or %.
- Velocity: Quantity/time
- Ratio (other than rate or %)

CHECKBOXES: MEASURE

- Rank
- Percentile
- X-ile: Quartile, Quintile or Decile
- Range: Mentions range or components (e.g., high, low)
- RR < 2
- RR > 3

CHECKBOXES: COMPARE

- CompQualTive: Comparison w/o numbers. E.g., more.
- CompQuanTive: Comparison w/numbers (e.g., 8% more)
- CompQuanRatio: Compares two ratios: %, rates.
- Attrib%: Percentage of cases attributed to factor
- Attrib#: Number of cases attributed to factor.
- ExplainBy%: Percent of variance explained by model.

CHECKBOXES: INFERENCE

- UnlikelyDueToChance: Used this phrase or idea.
- SampleUsed: Check yes, if sample is stated or likely.
- SampleRandom: Mentions use of a random sample.
- Margin Error: Mentions size of 95% margin of error.
- P-value: Mentions p-value.
- StatInsig?: Says result is not significant

TEXTBOXES (Middle or Bottom of form)

- SampleSize: Size of sample if given.
- FileName of PDF: Usually publication date and title.
- Owner: Organization publishing the article.
- URL: Source (may no longer be available on the web).

APPENDIX B: DATA SUMMARY

Table 8: Counts Overall and by Outcome Data Type
1=Multinomial, 2 = Binary, 3=Ordinal, 4=Discrete, 5=Continuous

Outcomes	ALL	0	1	2	3	4	5
OutcomeData	160	8	2	118	5	5	22
CauseAssert	62	1	2	45	1	2	11
CauseImply	37	0	0	29	1	1	6
AsmblPrsnt	43	5	2	24	4	2	6
Rev Plausible	7	0	0	6	0	0	1
FactorChngbl	91	1	1	67	1	3	18
OutcomeRptbl	59	0	1	42	3	3	10
Controlled	41	1	1	29	1	0	9
Longitudinal	51	2	0	37	1	2	9
Cohort	50	1	1	37	1	2	8
Manipulation	21	1	0	13	0	0	7
RandomAssign	4	0	0	4	0	0	0
ControlForModel	15	0	0	12	0	0	3
ControlForSelect	17	0	1	13	0	0	3
CnfndPlaus	67	2	1	55	2	2	5
AssmbPlaus	27	3	2	14	3	2	3
RndmPlaus	15	1	0	11	0	1	2
SubjBlindPoss	14	0	0	9	0	0	5
SubjectBlinded	7	0	0	5	0	0	2
ResBlindPoss	41	0	1	31	2	1	6
BiasPlaus	18	2	1	12	1	2	0
MultipleStudies	30	1	1	25	0	0	3
DataBinary	137	5	2	117	1	1	11
DataMulti	6	0	2	3	0	0	1
DataOrdinal	16	1	0	7	5	1	2
DataDiscrete	9	1	0	2	1	5	0
DataCont	49	2	0	22	1	3	21
Count	17	2	0	12	0	2	1
Measure	31	1	0	11	2	2	15
GrmPercent	86	5	1	70	2	3	5
GrmPercentage	2	0	0	2	0	0	0
GrmRate	50	3	1	39	1	2	4
GrmrChance	37	0	1	36	0	0	0
Incidence	7	0	0	5	0	1	1
Prevalence	81	3	1	72	2	1	2
Percent	113	8	2	85	2	3	13
Velocity	5	1	0	3	0	0	1
Ratio	20	2	0	14	1	2	1
Percentile	1	0	0	1	0	0	0
X-ile	2	0	0	1	0	1	0
Range	4	1	0	2	0	1	0
SlopeQual	8	1	0	4	1	0	2
SlopeQuan	7	0	0	5	0	0	2
CompQualitative	31	0	0	19	3	2	7
CompQuantitative	104	2	1	81	2	3	15
CompQuanRatios	10	0	0	9	0	0	1
Attribute%	3	0	0	2	0	0	1
Attribute#	3	0	0	2	0	0	1
RR<2	17	0	1	16	0	0	0
RR>3	13	1	0	12	0	0	0
UnlikelyDueChance	16	0	0	13	0	0	3
SampleUsed	127	7	2	94	5	3	16
SampleRandom	3	1	0	2	0	0	0
MarginError	1	1	0	0	0	0	0
StatInsig?	10	0	0	8	0	0	2

Table 9: Column Percentages by Outcome Data Type
1=Multinomial, 2 = Binary, 3=Ordinal, 4=Discrete, 5=Continuous

Outcomes	ALL	0	1	2	3	4	5
OutcomeData	100%	100%	100%	100%	100%	100%	100%
CauseAssert	39%	13%	100%	38%	20%	40%	50%
CauseImply	23%	0%	0%	25%	20%	20%	27%
AsmblPrsnt	27%	63%	100%	20%	80%	40%	27%
Rev Plausible	4%	0%	0%	5%	0%	0%	5%
FactorChngbl	57%	13%	50%	57%	20%	60%	82%
OutcomeRptbl	37%	0%	50%	36%	60%	60%	45%
Controlled	26%	13%	50%	25%	20%	0%	41%
Longitudinal	32%	25%	0%	31%	20%	40%	41%
Cohort	31%	13%	50%	31%	20%	40%	36%
Manipulation	13%	13%	0%	11%	0%	0%	32%
RandomAssign	3%	0%	0%	3%	0%	0%	0%
ControlForModel	9%	0%	0%	10%	0%	0%	14%
ControlForSelect	11%	0%	50%	11%	0%	0%	14%
CnfndPlaus	42%	25%	50%	47%	40%	40%	23%
AssmbPlaus	17%	38%	100%	12%	60%	40%	14%
RndmPlaus	9%	13%	0%	9%	0%	20%	9%
SubjBlindPoss	9%	0%	0%	8%	0%	0%	23%
SubjectBlinded	4%	0%	0%	4%	0%	0%	9%
ResBlindPoss	26%	0%	50%	26%	40%	20%	27%
BiasPlaus	11%	25%	50%	10%	20%	40%	0%
MultipleStudies	19%	13%	50%	21%	0%	0%	14%
DataBinary	86%	63%	100%	99%	20%	20%	50%
DataMulti	4%	0%	100%	3%	0%	0%	5%
DataOrdinal	10%	13%	0%	6%	100%	20%	9%
DataDiscrete	6%	13%	0%	2%	20%	100%	0%
DataCont	31%	25%	0%	19%	20%	60%	95%
Count	11%	25%	0%	10%	0%	40%	5%
Measure	19%	13%	0%	9%	40%	40%	68%
GrmPercent	54%	63%	50%	59%	40%	60%	23%
GrmPercentage	1%	0%	0%	2%	0%	0%	0%
GrmRate	31%	38%	50%	33%	20%	40%	18%
GrmrChance	23%	0%	50%	31%	0%	0%	0%
Incidence	4%	0%	0%	4%	0%	20%	5%
Prevalence	51%	38%	50%	61%	40%	20%	9%
Percent	71%	100%	100%	72%	40%	60%	59%
Velocity	3%	13%	0%	3%	0%	0%	5%
Ratio	13%	25%	0%	12%	20%	40%	5%
Percentile	1%	0%	0%	1%	0%	0%	0%
X-ile	1%	0%	0%	1%	0%	20%	0%
Range	3%	13%	0%	2%	0%	20%	0%
SlopeQual	5%	13%	0%	3%	20%	0%	9%
SlopeQuan	4%	0%	0%	4%	0%	0%	9%
CompQualitative	19%	0%	0%	16%	60%	40%	32%
CompQuantitative	65%	25%	50%	69%	40%	60%	68%
CompQuanRatios	6%	0%	0%	8%	0%	0%	5%
Attribute%	2%	0%	0%	2%	0%	0%	5%
Attribute#	2%	0%	0%	2%	0%	0%	5%
RR<2	11%	0%	50%	14%	0%	0%	0%
RR>3	8%	13%	0%	10%	0%	0%	0%
UnlikelyDueChance	10%	0%	0%	11%	0%	0%	14%
SampleUsed	79%	88%	100%	80%	100%	60%	73%
SampleRandom	2%	13%	0%	2%	0%	0%	0%
MarginError	1%	13%	0%	0%	0%	0%	0%
StatInsig?	6%	0%	0%	7%	0%	0%	9%

APPENDIX C: DATA MATCHING PROCESS AND RESULTS

The original data source was 899 pdf files: the content of number-based news stories. The content of the article involved copying the text of a PDF into Notepad and saving it as ASCIII text. This story text and the associated filename were imported into an Access database with two fields per record: FileName and Story (a memo field).

Counts of matches were obtained by programmatically matching 231 keywords with the content of the articles using the Access "Like" command. To insure the search text was not part of a larger word (e.g., 'cause' was not part of 'because'), the text was required to be preceded and followed by non-text characters: [~A-Z]. Here is the SQL for StudyTitle: IIf([FileName] Like "*study*",1,0). Here is the SQL for Study: IIf([Story] Like "*[!A-Z]study[!A-Z]*",1,0). This process gives an upper limit since the meaning of words varies with their context.

There were some special cases. The 'Associated' shown here specifically excludes 'Associated Press.' Associated: IIf([Story] Like "*[!A-Z]associated[!A-Z][!Press]*",1,0). A few words having multiple acceptable endings were noted with an asterisk at the end such as Outlier*: IIf([Story] Like "*[!A-Z]outlier*",1,0).

To eliminate the need for quotes around phrases, they are shown as an unbroken string. Thus, TimesAsMuchAs: was the name for this SQL query: IIf([Story] Like "*[!A-Z]times as much as[!A-Z]*",1,0).

These 103 terms had matches in at least two percent of the 899 articles:
More (90.0%), Study (75.0%), Percent (65.3%), Report (65.1%), Because (47.9%), Risk (47.9%), MoreThan (46.8%), Likely (42.6%), PercentOf (42.3%), Less (41.7%), RiskOf (34.7%), LikelyTo (33.1%), ErThan (32.1%), Results (27.3%), Average (27.1%), Rate (25.4%), MoreLikely (24.9%), Association (23.8%), Related (23.8%), Control (23.4%), Times (22.8%), Cause (21.9%), MoreLikelyTo (20.7%), Population (19.7%), Associated (19.6%), Effect (19.0%), Effects (18.8%), Significant (18.8%), Per (16.7%), Linked (16.1%), Survey (15.9%), StudyTitle (15.4%), Causes (14.1%), LessThan (12.8%), RateOf (12.3%), Percentage (12.0%), Link (11.6%), LessLikely (10.7%), Caused (10.2%), Factor (9.9%), % (9.8%), LessLikelyTo (9.2%), Relationship (9.1%), Result (8.9%), Account (8.5%), LikelyThan (8.5%), EffectsOf (8.2%), Sample (8.0%), %Of (7.9%), TheRate (7.9%), Mean (7.1%), MoreLikelyThan (6.8%), TimesMore (6.6%), Error (6.6%), Range (6.3%), Chance (6.2%), CausedBy (6.1%), PercentagePoints (5.9%), Sampling (5.8%), Share (5.7%), PercentageOf (5.3%), TheRateOf (5.3%), Incidence (5.3%), Prevalence (5.2%), ResultsOf (4.8%), ConfidenceLevel (4.8%), EffectOf (4.7%), ThePercentage (4.7%), %Confidence (4.4%), 95%Confiden* (4.4%), IncidenceOf (4.1%), AsLikelyTo (4.1%), Attributed (3.9%), Percentages (3.9%), Odds (3.9%), Likelihood (3.9%), PercentMore (3.9%), ThePercentageOf (3.8%), ChanceOf (3.8%), AccountFor (3.7%), IntoAccount (3.7%), PrevalenceOf (3.7%), RangeOf (3.6%), Relation (3.2%), Placebo (3.2%), LikelihoodOf (3.1%), AttributedTo (3.0%), ResultOf (2.9%), Ratio (2.9%), ResultIn (2.7%), RandomSample (2.7%), RandomlySampled (2.7%), Representative (2.7%), StatisticallySignificant (2.7%), PercentIncrease (2.6%), Probability (2.4%), LessLikelyThan (2.4%), Ranked (2.3%), ProbabilityOf (2.3%), LikelyAs (2.3%), ResultsIn (2.2%), RandomlyAssigned (2.1%), AsLikelyAs (2.0%),

The following 80 terms had matches but matched in less than two percent of the 899 articles:
Experiment (1.8%), OddsOf (1.8%), TimesAs (1.7%), AccountsFor (1.6%), Median (1.6%), Resulted (1.4%), ControlGroup (1.4%), ShareOf (1.4%), ReportTitle (1.4%), ChanceTo (1.3%), RatioOf (1.3%), NumMoreThan (1.3%), SurveyTitle (1.2%), Attribute (1.2%), Randomized (1.1%), Fraction (1.1%), Relate (1.0%), ControlOf (1.0%), Attributable (1.0%), Attributes (1.0%), AttributableTo (1.0%), ResultedIn (0.9%), ClinicalTrial (0.9%), Ranks (0.9%), FractionOf (0.9%), Causal (0.8%), Correlated (0.8%), ControlFor (0.8%), RangeFrom (0.8%), PercentChance (0.8%), RiskThat (0.8%), TimesLess (0.8%), TakenIntoAccount (0.7%), ControllingFor (0.7%), Bias (0.7%), %More (0.7%), Percentile (0.6%), Correlate (0.4%), TakeIntoAccount (0.4%), TakingIntoAccount (0.4%), ControlledStudy (0.4%), Longitudinal (0.4%), Rank (0.4%), Standardize* (0.4%), LikelihoodThat (0.4%), Sampled (0.4%), MarginOfError (0.4%), AccountOf (0.3%), TakesIntoAccount (0.3%), PlaceboEffect (0.3%), Skewed (0.3%), Regress* (0.3%), ChanceThat (0.3%), OddsThat (0.3%), PercentDecrease (0.3%), TimesAsMuch (0.3%), Confounding (0.2%), ControlledFor (0.2%), ObservationalStudy (0.2%), RandomlyAssign (0.2%), Biased (0.2%), Mode (0.2%), StdDev (0.2%), RelativeRisk (0.2%), OddsRatio (0.2%), Normalize* (0.2%), PercentMoreThan (0.2%), PercentLessThan (0.2%), TimesMoreThan (0.2%), NotAsignificant (0.2%), Percentiles (0.1%), Quartile (0.1%), PercentagePoint (0.1%), Outlier* (0.1%), ChanceFor (0.1%), %Less (0.1%), PercentLess (0.1%), %Decrease (0.1%), TimesLessThan (0.1%), DueToChance (0.1%), InSignificant (0.1%),

The following 48 terms had no matches in these 899 articles:
Effected (0.0%), Confounded (0.0%), Confounder (0.0%), Spurious (0.0%), ExposureGroup (0.0%), TreatmentGroup (0.0%), RandomAssignment (0.0%), CrossSection (0.0%), BlindSingle (0.0%), BlindDouble (0.0%), Quintile (0.0%), Symmetric (0.0%), WeightedAverage (0.0%), CoefVar (0.0%), EffectSize (0.0%), Zscore (0.0%), GiniCoef (0.0%), PercentAttributableTo (0.0%), PercentAttributedTo (0.0%), PercentShare (0.0%), PercentFraction (0.0%), %Accuracy (0.0%), PercentAccuracy (0.0%), PercentConfidence (0.0%), PercentageThat (0.0%), PercentageWho (0.0%), ThePercentageThat (0.0%), ThePercentageWho (0.0%), %Chance (0.0%), OddsTo (0.0%), PercentProbability (0.0%), ProbabilityThat (0.0%), %MoreThan (0.0%), NumLessThan (0.0%), %LessThan (0.0%), %Off (0.0%), %Increase (0.0%), PercentOff (0.0%), TimesAsMuchAs (0.0%), UnlikelyDueToChance As (0.0%), %MarginOfError (0.0%), 95%MarginOfError (0.0%), PercentConfident (0.0%), %Confident (0.0%), ConfidenceInterval (0.0%), StatisticalSignificance (0.0%), NotSignificant (0.0%),