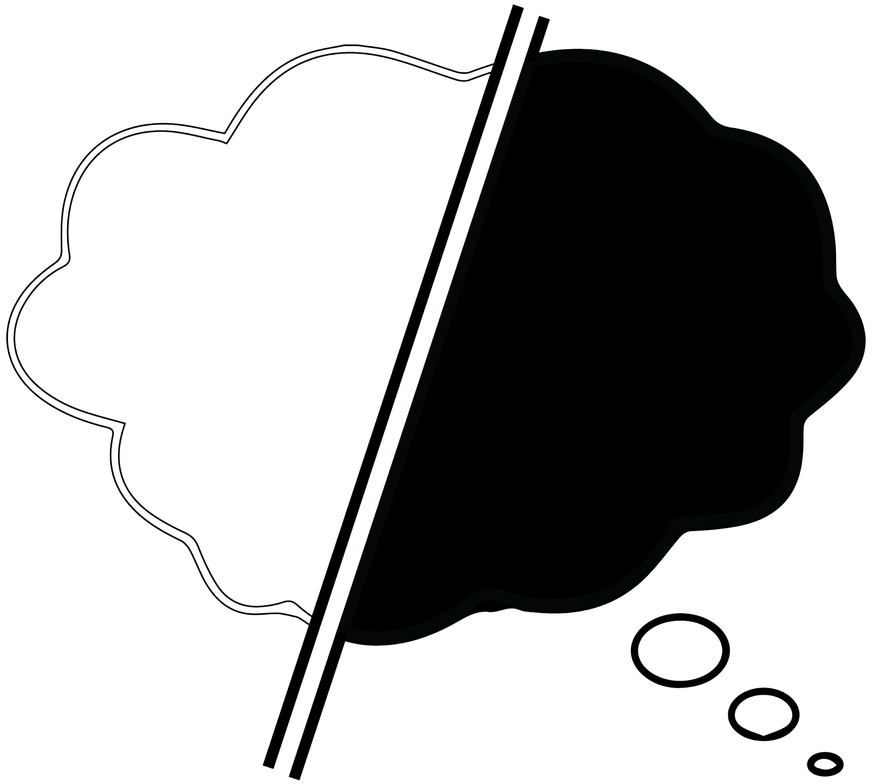
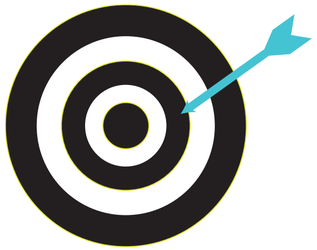
The Continued Prevalence of Dichotomous Inferences at CHI

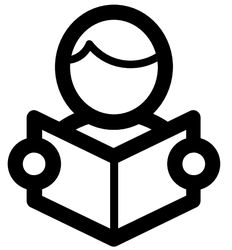


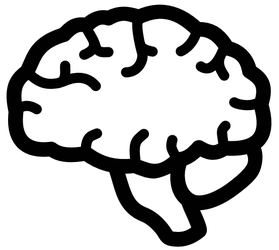
LINKÖPING UNIVERSITY

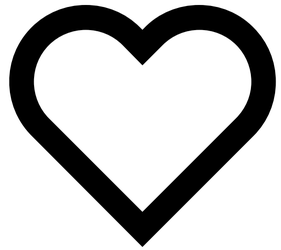
Lonni Besançon, Pierre Dragicevic

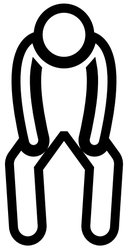
Aviz

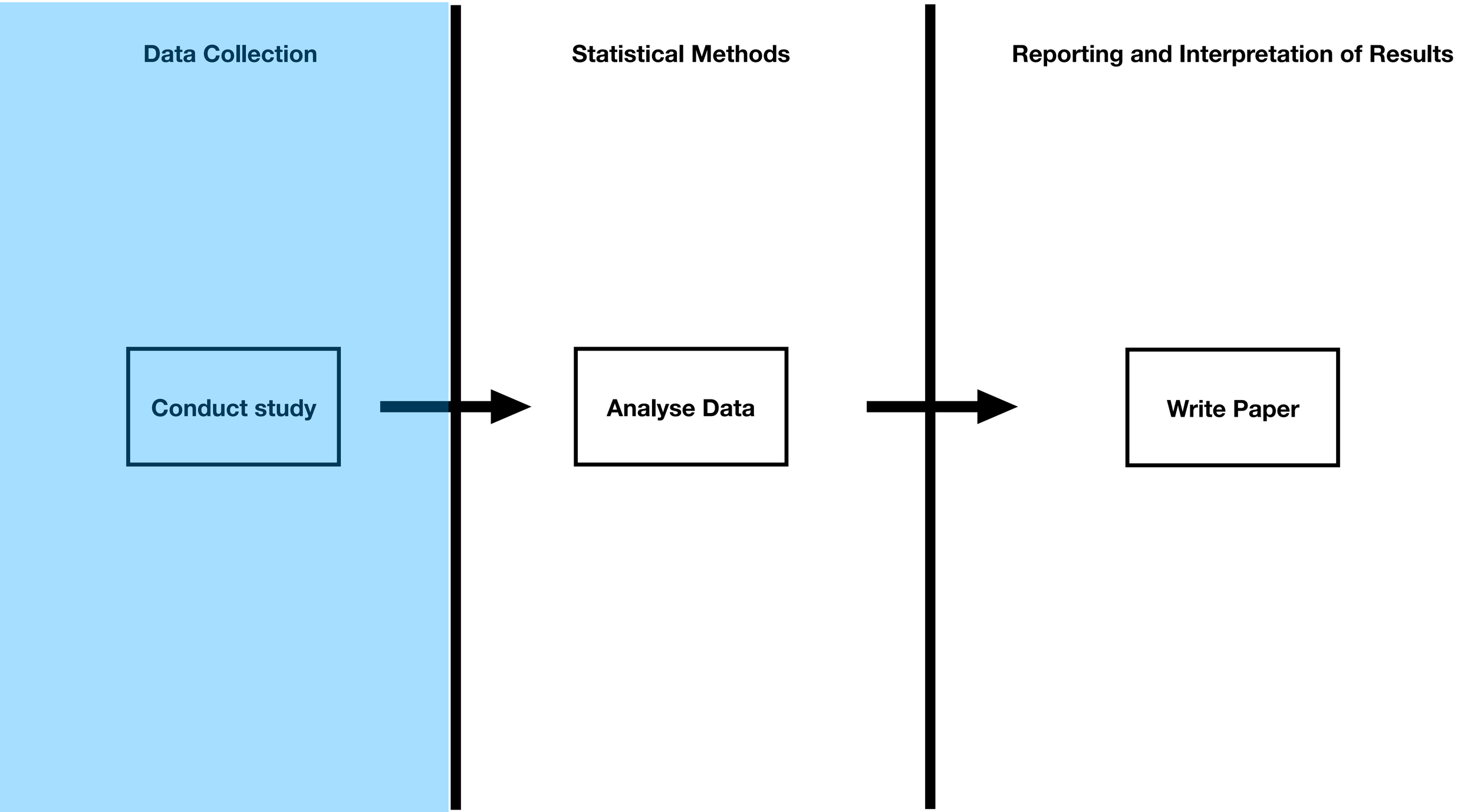


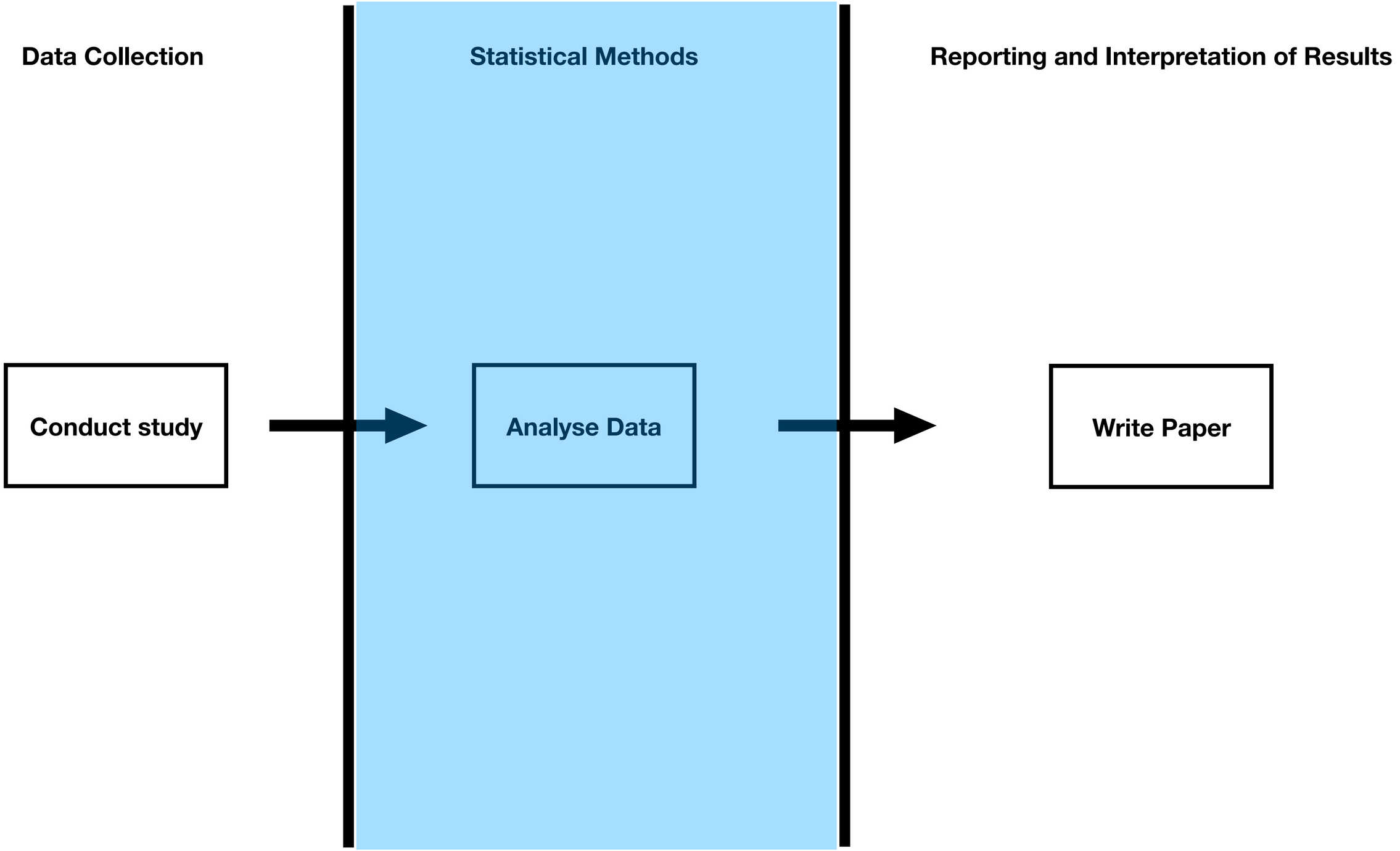


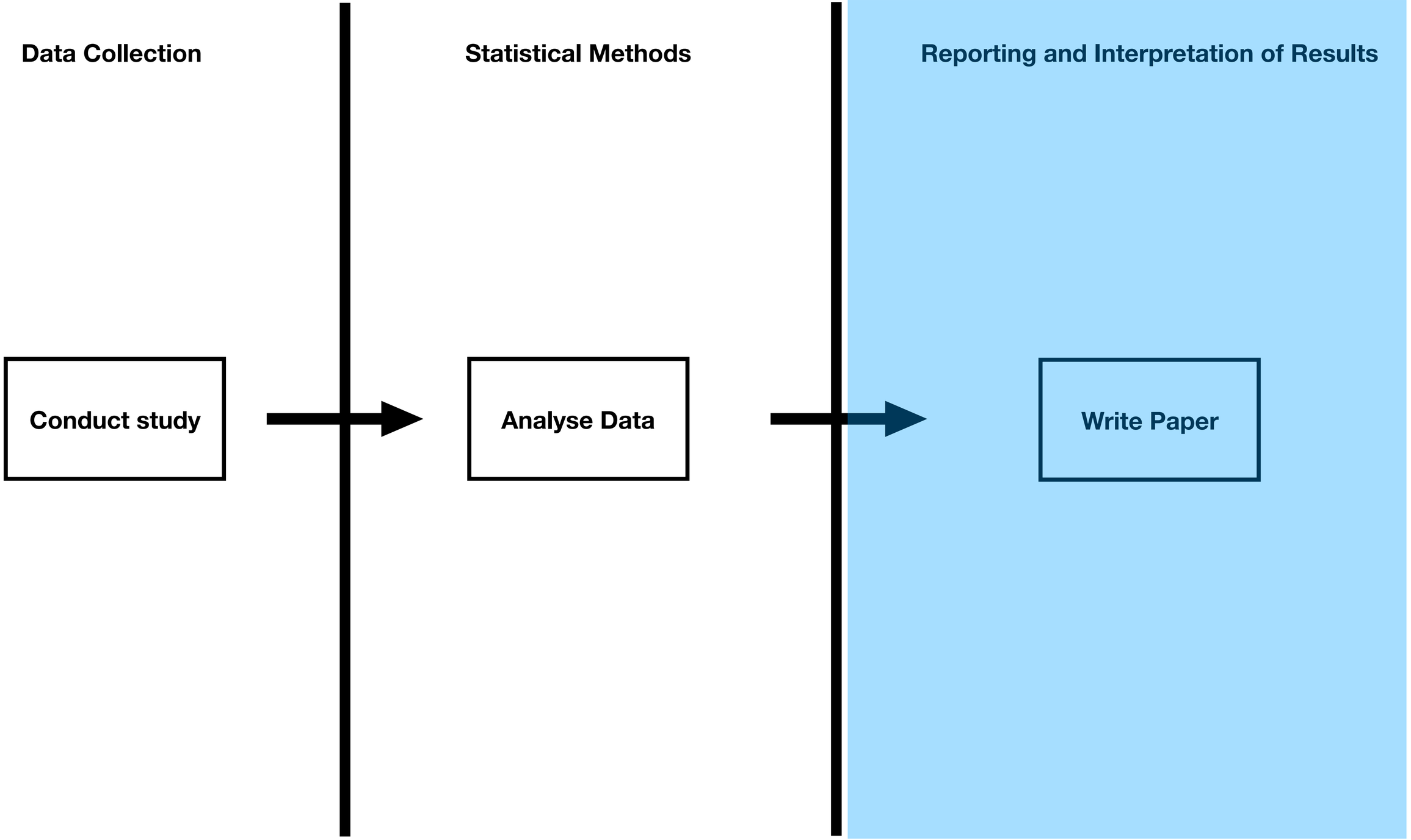


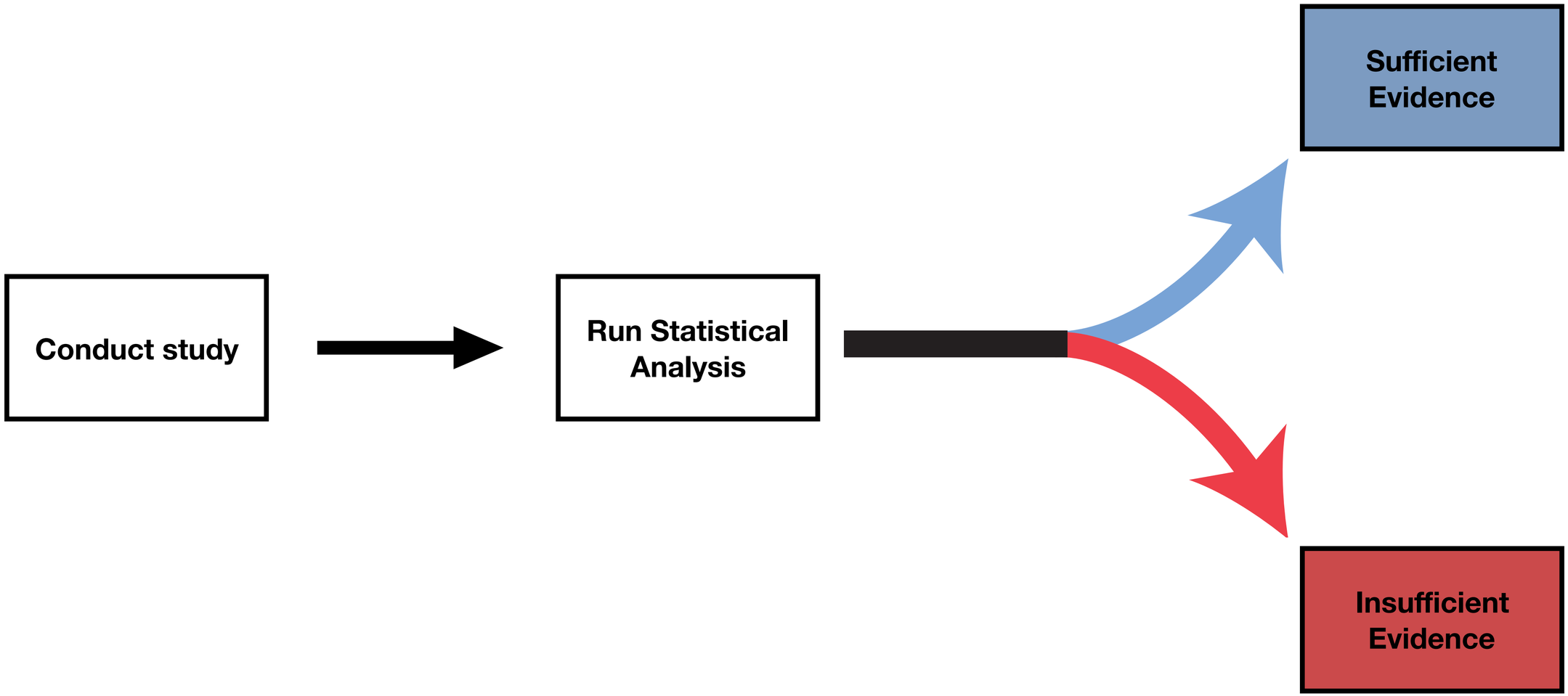


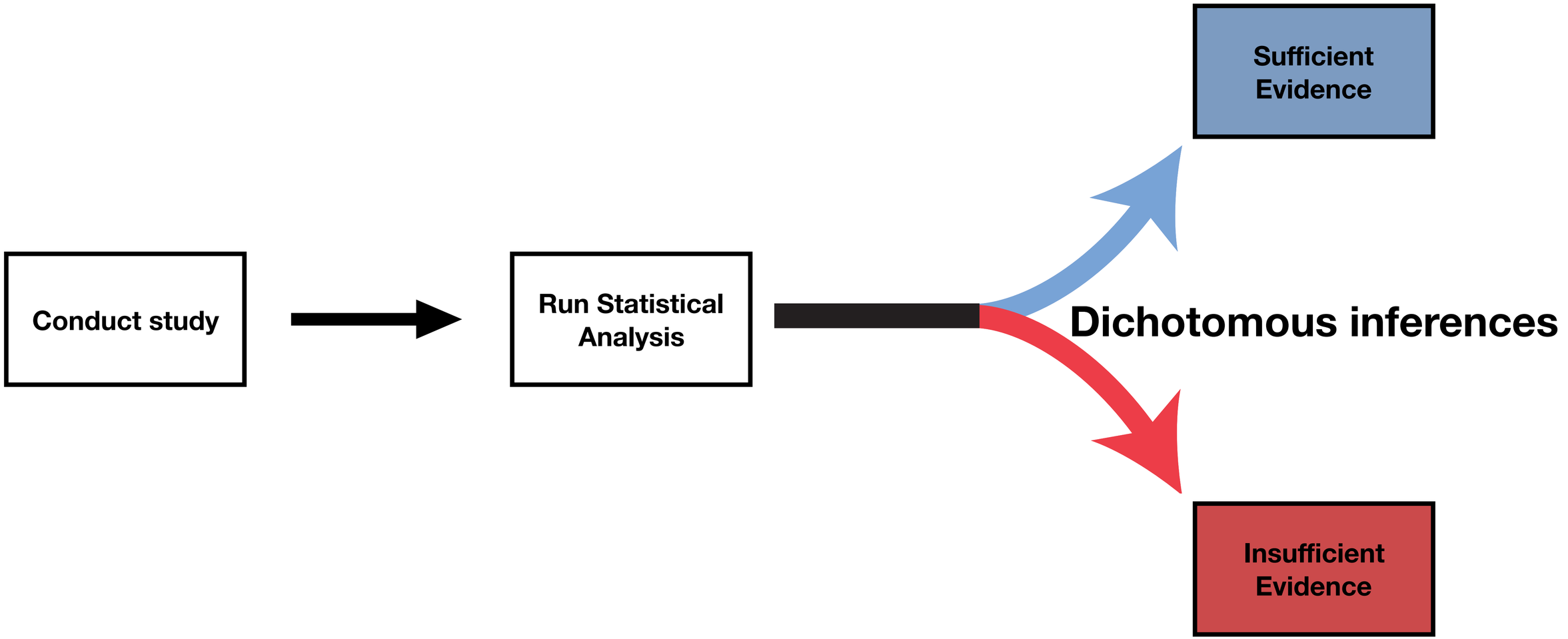


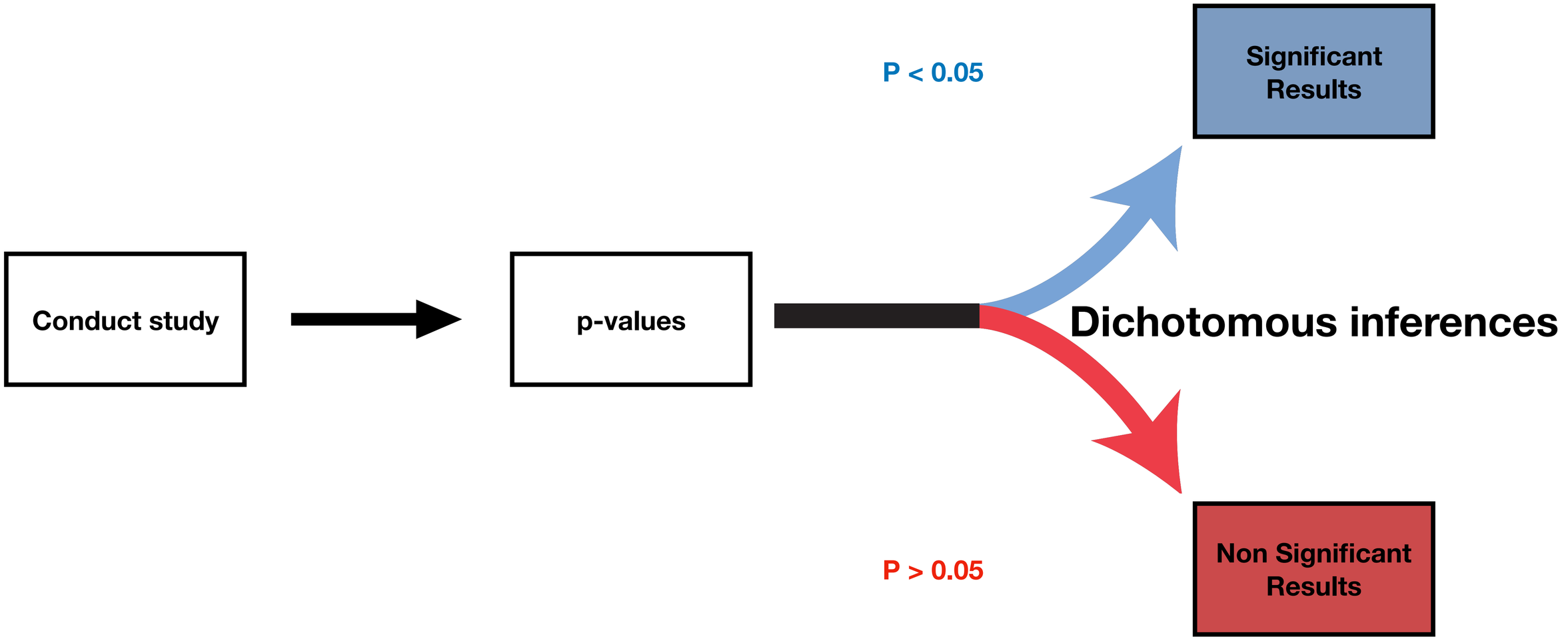

Data Collection
Statistical Methods
Reporting and Interpretation of Results
Conduct study
Analyse Data
Write Paper

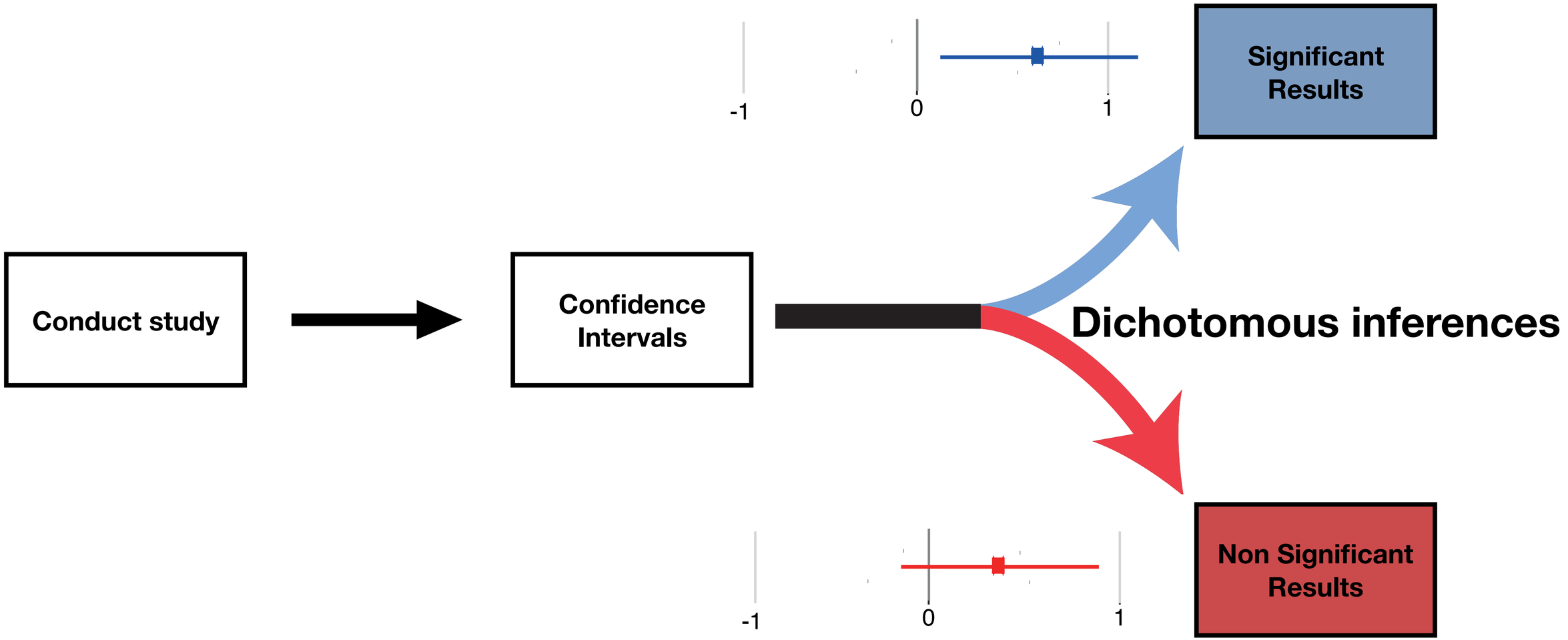

Data Collection
Statistical Methods
Reporting and Interpretation of Results
Conduct study
Analyse Data
Write Paper

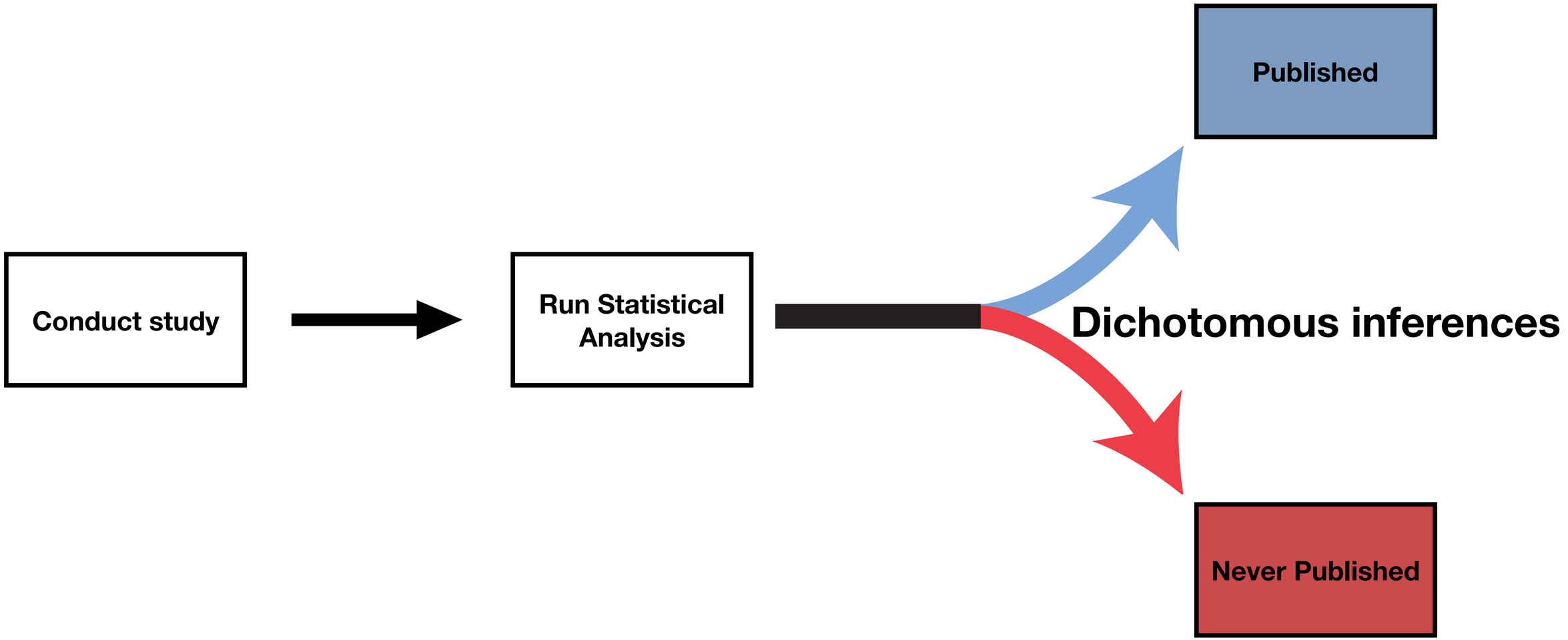

Data Collection
Statistical Methods
Reporting and Interpretation of Results
Conduct study
Analyse Data
Write Paper


Conduct study
Run Statistical Analysis
Sufficient Evidence
Insufficient Evidence

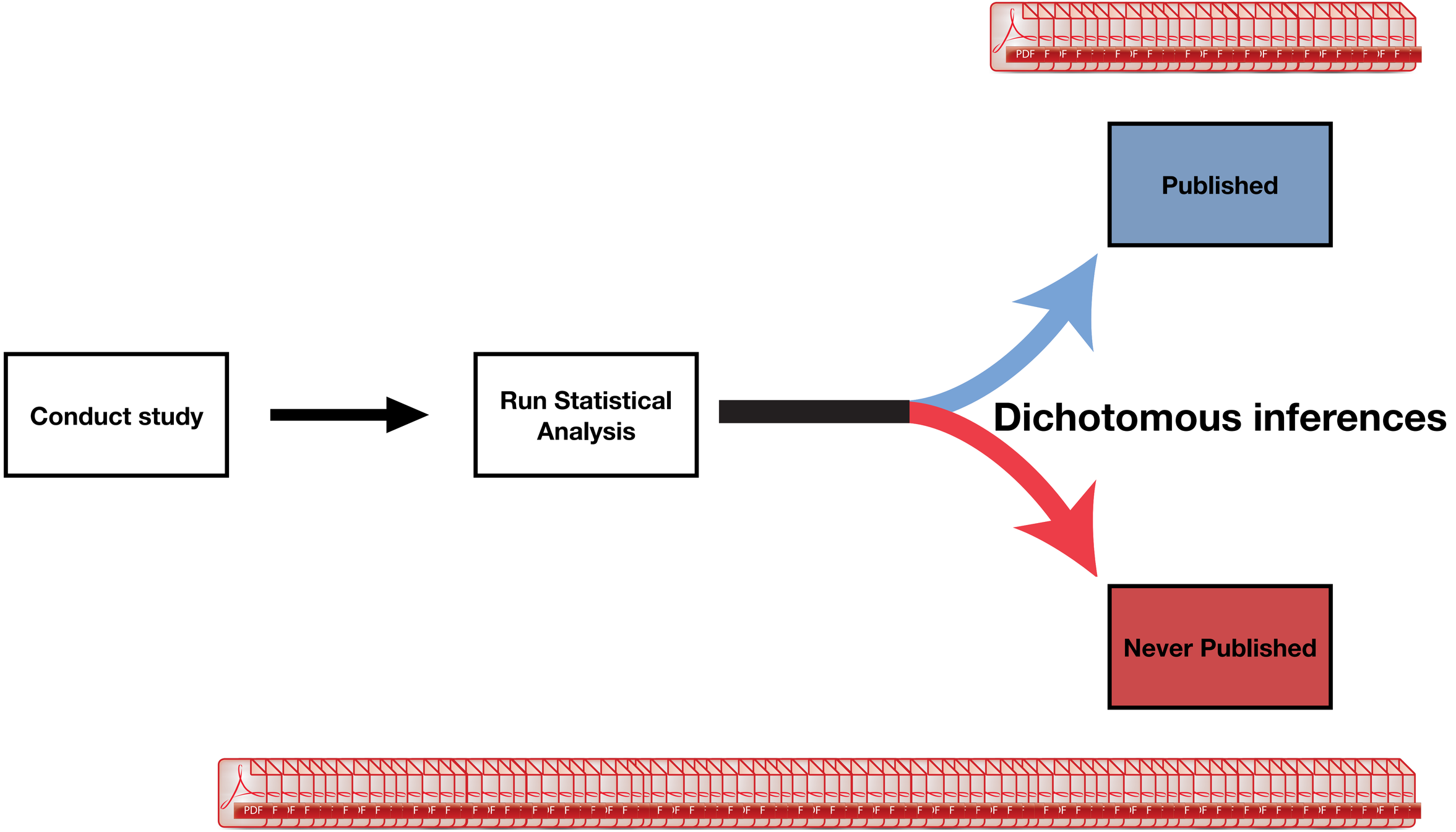

Conduct study
Run Statistical Analysis
Dichotomous inferences
Sufficient Evidence
Insufficient Evidence

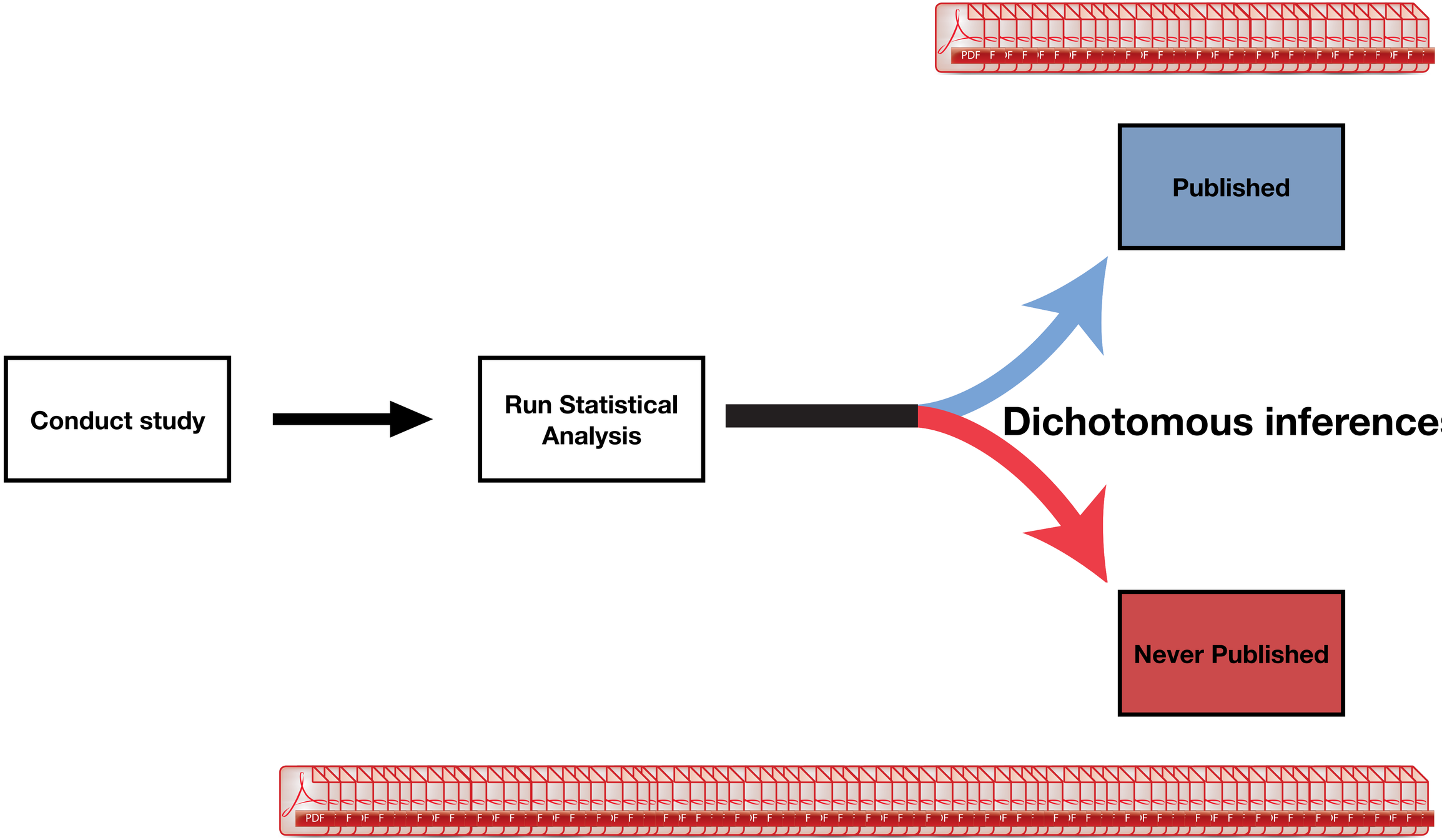

Conduct study
p-values
P < 0.05
Significant Results
Dichotomous inferences
P > 0.05
Non Significant Results

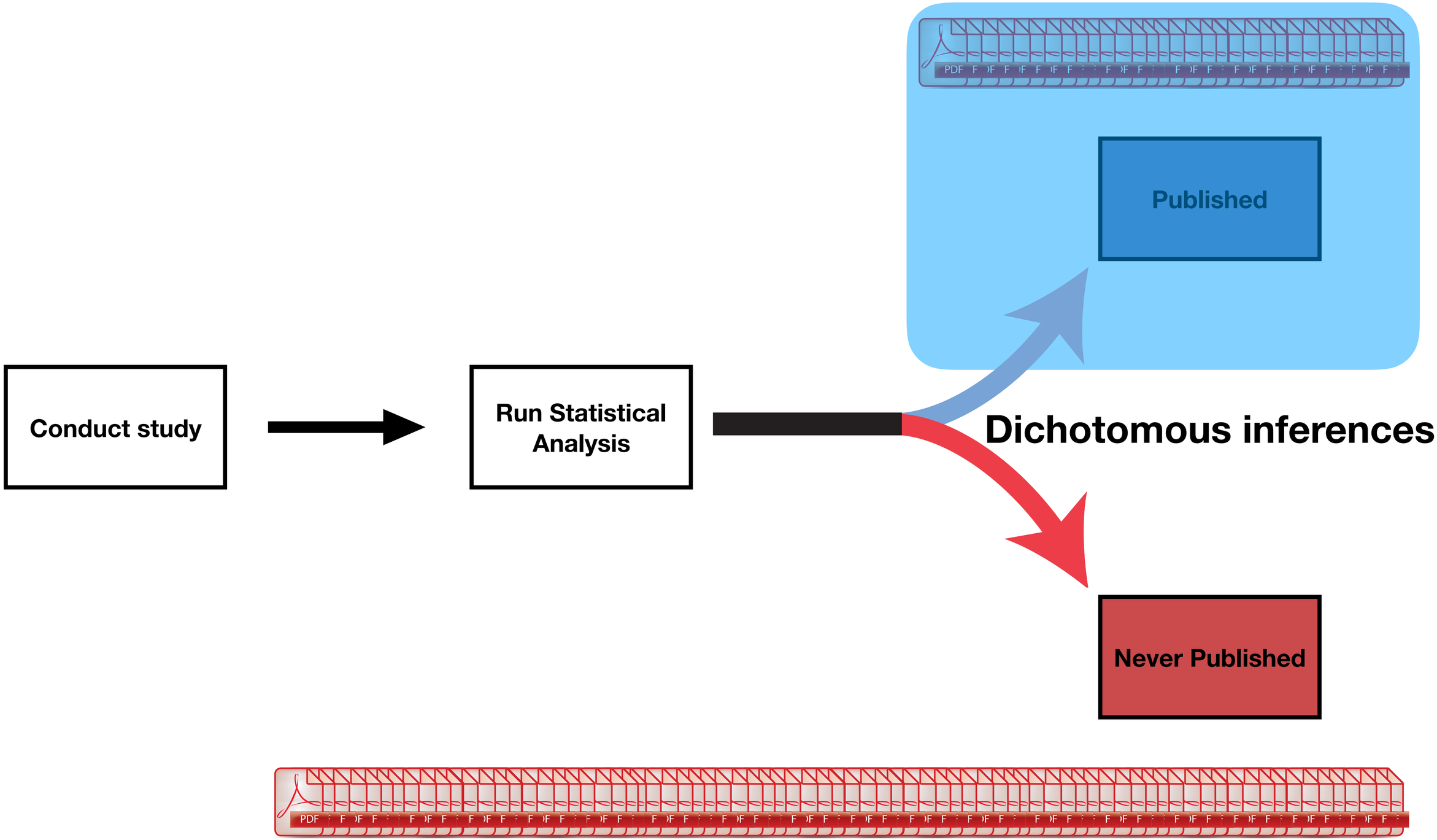

Conduct study
Confidence Intervals
-1
0
1
Significant Results
Dichotomous inferences
-1
0
1
Non Significant Results


Conduct study
Run Statistical Analysis
Dichotomous inferences
Published
Never Published

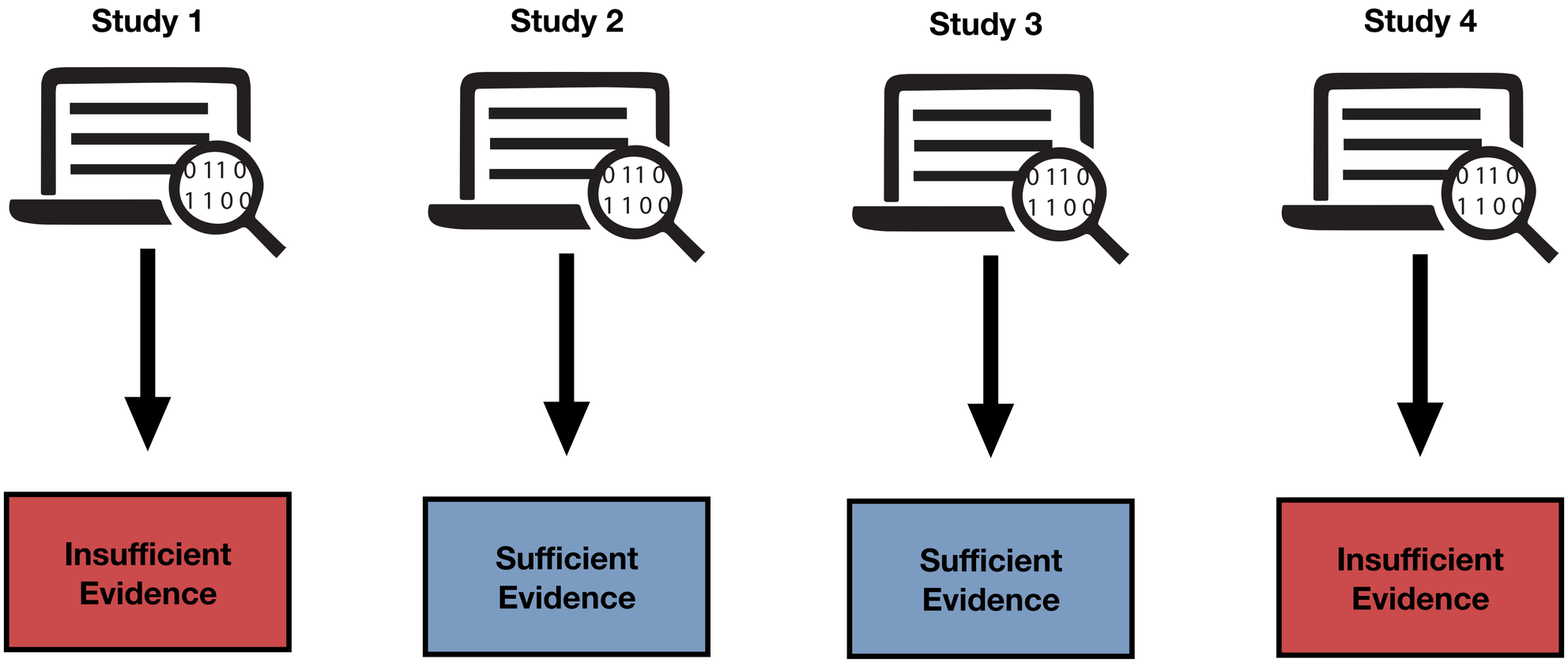
...and this is where we put the non-significant results.


Conduct study
Run Statistical Analysis
Dichotomous inferences
Published
Never Published

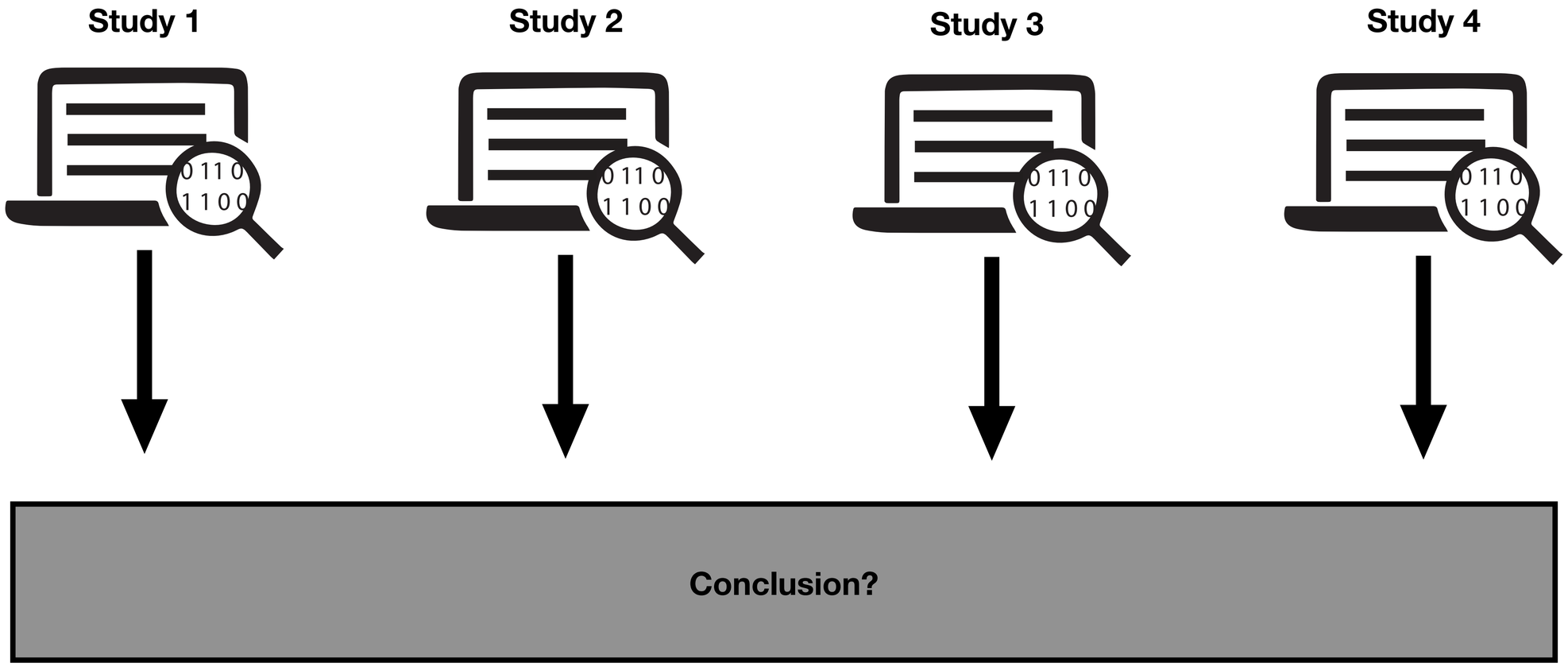

Conduct study
Run Statistical Analysis
Dichotomous inferences
Published
Never Published


Conduct study
Run Statistical Analysis
Published
Dichotomous inferences
Never Published

Is drug Z efficient against disease D?

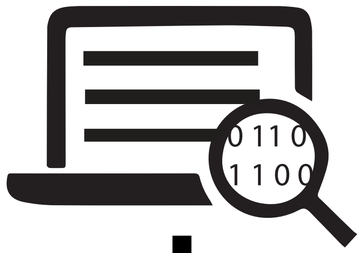


Is drug Z efficient against disease D?



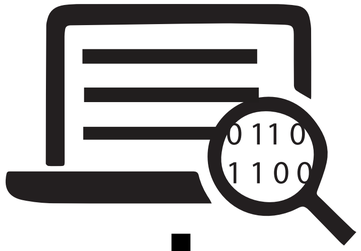
Is drug Z efficient against disease D?

Study 1



p = 0.075
CI= […;…]

Study 2

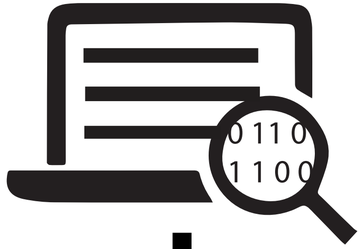


p = 0.012
CI= […;…]

Study 3

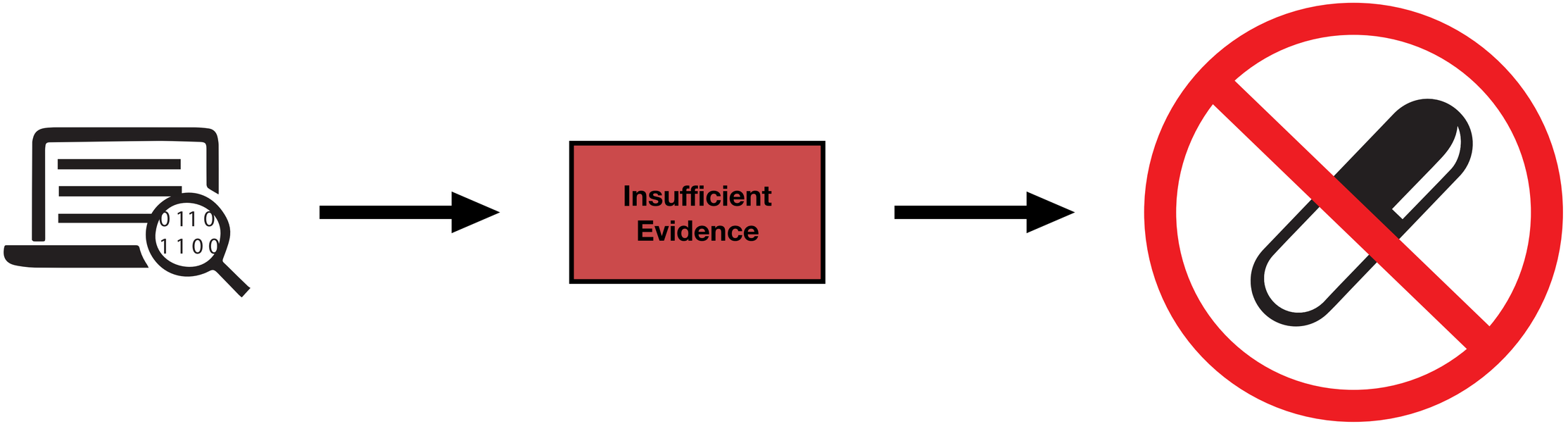
p = 0.031
CI= […;…]

Study 4

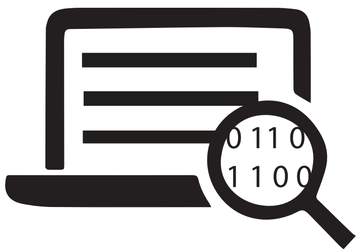


p = 0.062
CI= […;…]

Is drug Z efficient against disease D?



Does drug Z have secondary effects on patients?



Insufficient Evidence

PEER-REVIEWED

The earth is flat ($p > 0.0$ thresholds and the crisis research

Literature review | Science Policy | Statistics

We're using a common statistical test all wrong. Statisticians want to fix that.

After reading too many papers that either are not reproducible or contain statistical errors (or both),

ASA AMERICAN STATISTICAL ASSOCIATION Promoting the Practice and Profession of Statistics

Original Articles

Life After NHST: How to Describe Your Data Without "*p*-ing" Everywhere

Jeffrey C. Valentine, Ariel M. Aloe & Timothy S. Lau

Pages 260-273 | Published online: 04 Aug 2015

Download citation | https://doi.org/10.1080/01973533.2015.1060240 | Check for updates

The Journal of Socio-Economics

er.com/locate/econbase

Pierre Dragicevic Inria
Pierre Dragicevic Inria
Pierre Dragicevic Inria
Pierre Dragicevic Inria
Fanny Chevalier Inria

Pierre Dragicevic Inria
Pierre Dragicevic Inria
Pierre Dragicevic Inria
Pierre Dragicevic Inria
Fanny Chevalier Inria

Mindless statistics

Gerd Gigerenzer*

Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

COMMENT · 20 MARCH 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Valentin Amrhein ✉, Sander Greenland & Blake McShane

https://www.nature.com/articles/d41586-019-00857-9

How do we fare in HCI?

Are we subject to dichotomous inferences in our research papers?

What methods do we use and does it influence how dichotomous we are?

Did the numerous literature on dichotomous interpretation affect us over the years?


CHI
2018


CHI 2017
EXPLORE INNOVATE INSPIRE


2016 #chi4good


CHI 2015
CROSSINGS
SEOUL·KOREA


CHI 2014
One of a CHInd


CHI2013
changing perspectives

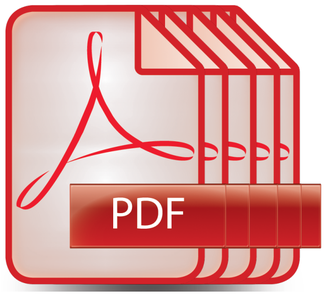

CHI 2012
it's the experience!
AUSTIN • TEXAS
MAY 5-10, 2012


2011
CHI
VANCOUVER, BC
MAY 7–12, 2011

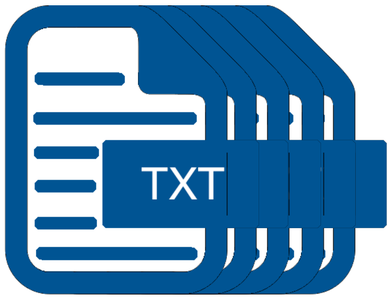

CHI 2010
we are HCI
10 - 15 April 2010
Atlanta, GA, USA

PDFs from CHI Papers 2010 to 2018

TXT version of the papers



x 4234



What do we report at CHI?

p-value inequalities:

“p <”, “p<”, “p >”, “p>”

p-value exact:

“p =”, “p=”

Confidence intervals:

“confidence interval”, “%ci”, “% ci”

p-value inequalities:

“p <”, “p<”, “p >”, “p>”

Inequalities only

p-value exact:

“p =”, “p=”

Exact only

(p < X with X = 0.01 accepted)

Confidence intervals:

“confidence interval”, “%ci”, “% ci”

p-value inequalities:

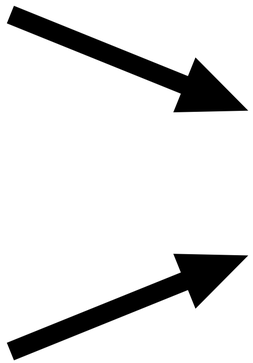
“p <”, “p<”, “p >”, “p>”

p-value exact:

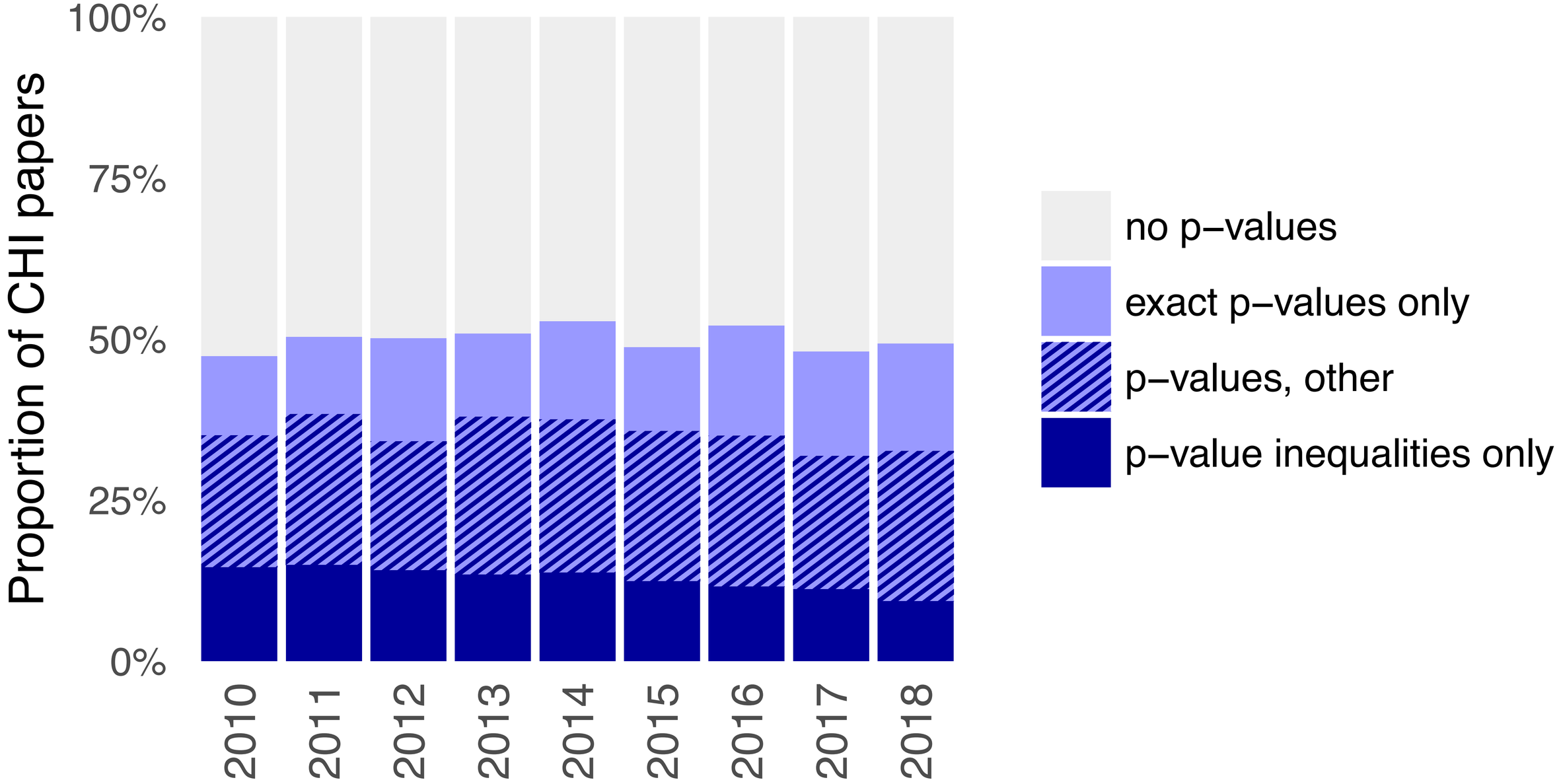
“p =”, “p=”

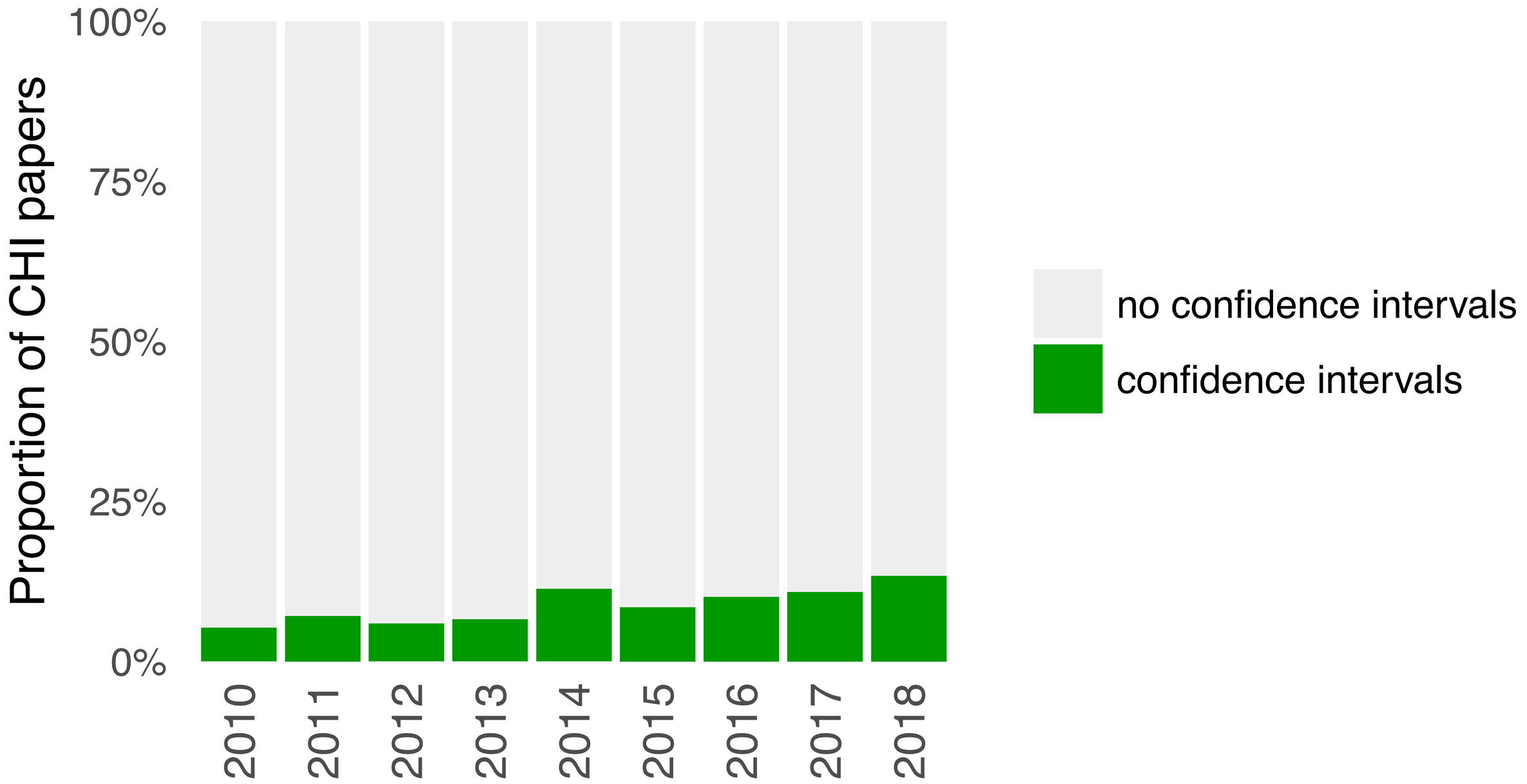
Confidence intervals:

“confidence interval”, “%ci”, “% ci”

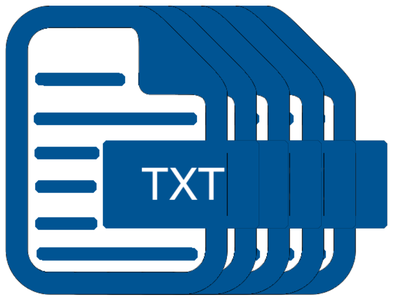


Ambiguous


Proportion of CHI papers
100%
75%
50%
25%
0%
2010
2011
2012
2013
2014
2015
2016
2017
2018
no p–values
exact p–values only
p–values, other
p–value inequalities only


Proportion of CHI papers
100%
75%
50%
25%
0%
2010
2011
2012
2013
2014
2015
2016
2017
2018
no confidence intervals
confidence intervals

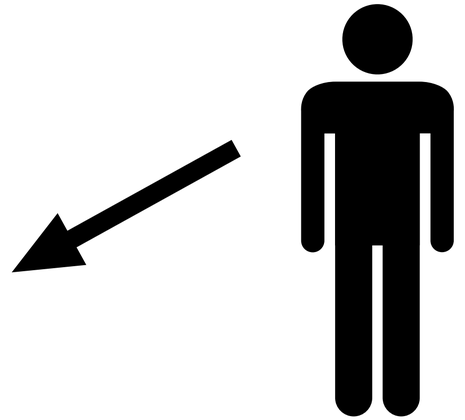
Are we dichotomous in our result interpretations?

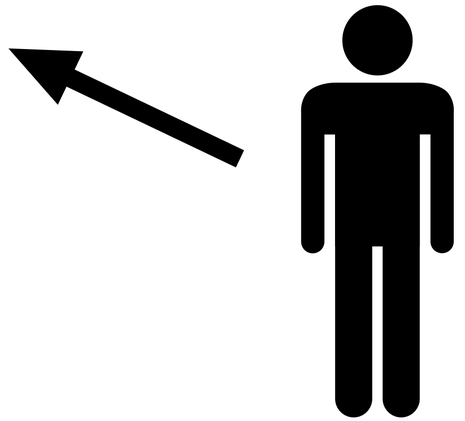


python

List of trigrams containing “significant” or “significantly”

Trigram	Occurences
No significant difference	1234
With significant others	678
A significant contribution	25





List of trigrams containing “significant” or “significantly”

Trigram	Occurences
No significant difference	1234
With significant others	678
A significant contribution	25
Found significant differences	13
Their significant others	10
A significant body	9
A significant interaction	8
A signifiant effect	2

No significant difference	1234
With significant others	678
A significant contribution	25
Found significant differences	13
Their significant others	10
A significant body	9
A significant interaction	8
A signifiant effect	2
A significant paper	1
The significant contribution	1

x 10,334 trigrams

No significant difference	1234
With significant others	678
A significant contribution	25
Found significant differences	13
Their significant others	10
A significant body	9
A significant interaction	8
A signifiant effect	2
A significant paper	1
The significant contribution	1

x 10,334 trigrams

No significant difference	1234
With significant others	678
A significant contribution	25
Found significant differences	13
Their significant others	10
A significant body	9
A significant interaction	8

x 1250 trigrams

List of trigrams containing “significant” or “significantly”

Trigram	Occurences	Coding Pierre	Coding Lonni
No significant difference	1234	True	True
With significant others	678	False	False
A significant contribution	25	False	False
Found significant differences	13	True	True
Their significant others	10	False	False
A significant body	9	False	False
A significant interaction	8	False	True

List of trigrams containing “significant” or “significantly”

Trigram	Occurences	Coding Pierre	Coding Lonni
No significant difference	1234	True	True
With significant others	678	False	False
A significant contribution	25	False	False
Found significant differences	13	True	True
Their significant others	10	False	False
A significant body	9	False	False
A significant interaction	8	False	True

List of trigrams containing “significant” or “significantly”

Trigram	Occurences	Coding Pierre	Coding Lonni	Likely Significance Language
No significant difference	1234	True	True	
With significant others	678	False	False	
A significant contribution	25	False	False	
Found significant differences	13	True	True	
Their significant others	10	False	False	
A significant body	9	False	False	
A significant interaction	8	False	True	

List of trigrams containing “significant” or “significantly”

Trigram	Occurences	Coding Pierre	Coding Lonni	Likely Significance Language
No significant difference	1234	True	True	True
With significant others	678	False	False	
A significant contribution	25	False	False	
Found significant differences	13	True	True	True
Their significant others	10	False	False	
A significant body	9	False	False	
A significant interaction	8	False	True	

List of trigrams containing “significant” or “significantly”

Trigram	Occurences	Coding Pierre	Coding Lonni	Likely Significance Language
No significant difference	1234	True	True	True
With significant others	678	False	False	
A significant contribution	25	False	False	
Found significant differences	13	True	True	True
Their significant others	10	False	False	
A significant body	9	False	False	
A significant interaction	8	False	True	

List of trigrams containing “significant” or “significantly”

Trigram	Occurences	Coding Pierre	Coding Lonni	Likely Significance Language
No significant difference	1234	True	True	True
With significant others	678	False	False	False
A significant contribution	25	False	False	False
Found significant differences	13	True	True	True
Their significant others	10	False	False	False
A significant body	9	False	False	False
A significant interaction	8	False	True	

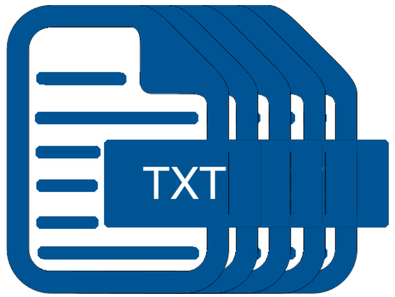
List of trigrams containing “significant” or “significantly”

Trigram	Occurences	Coding Pierre	Coding Lonni	Likely Significance Language
No significant difference	1234	True	True	True
With significant others	678	False	False	False
A significant contribution	25	False	False	False
Found significant differences	13	True	True	True
Their significant others	10	False	False	False
A significant body	9	False	False	False
A significant interaction	8	False	True	

List of trigrams containing “significant” or “significantly”

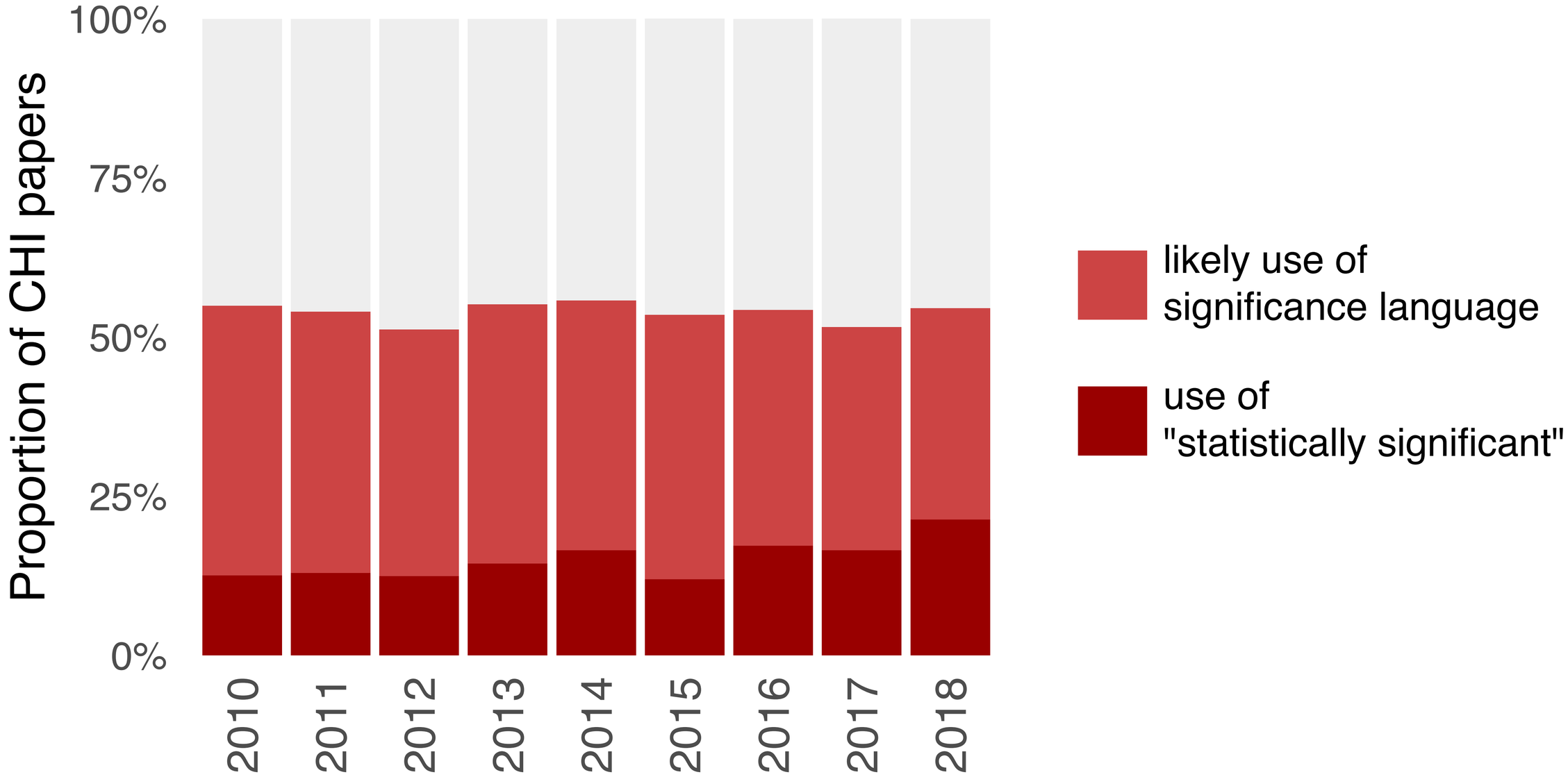
Trigram	Occurences	Coding Pierre	Coding Lonni	Likely Significance Language
No significant difference	1234	True	True	True
With significant others	678	False	False	False
A significant contribution	25	False	False	False
Found significant differences	13	True	True	True
Their significant others	10	False	False	False
A significant body	9	False	False	False
A significant interaction	8	False	True	False

List of Likely significance language trigrams

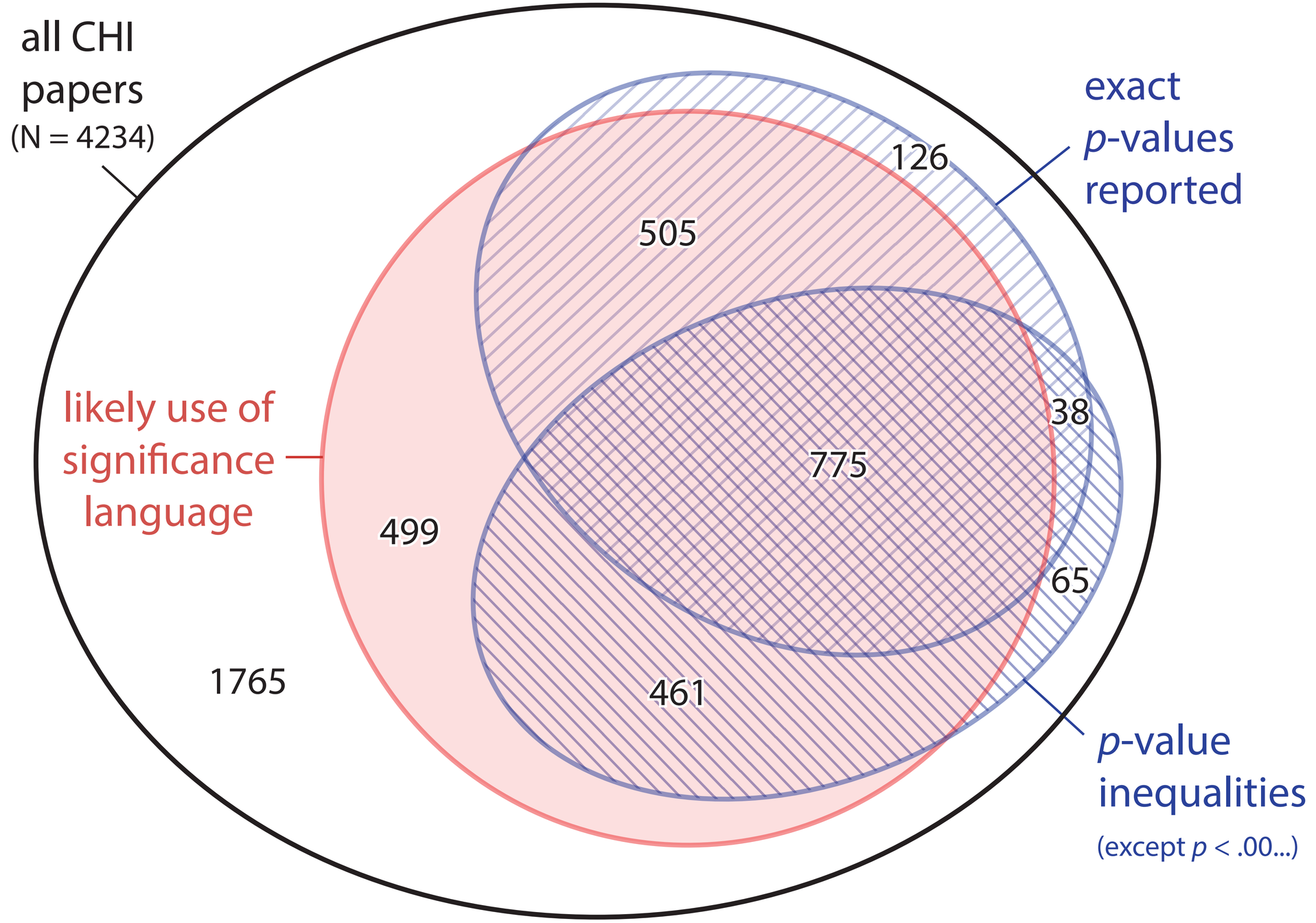


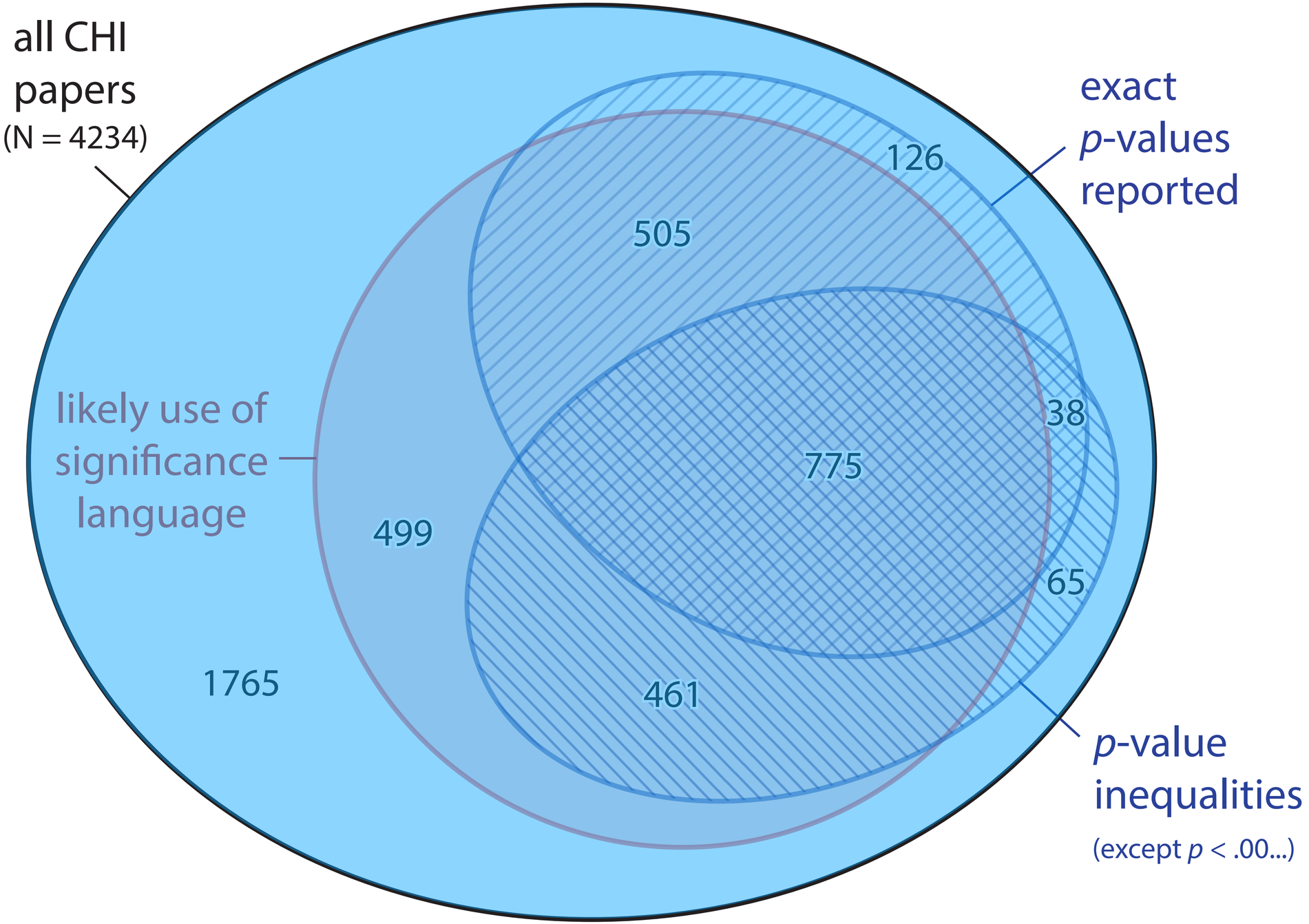
List of papers containing likely significance language

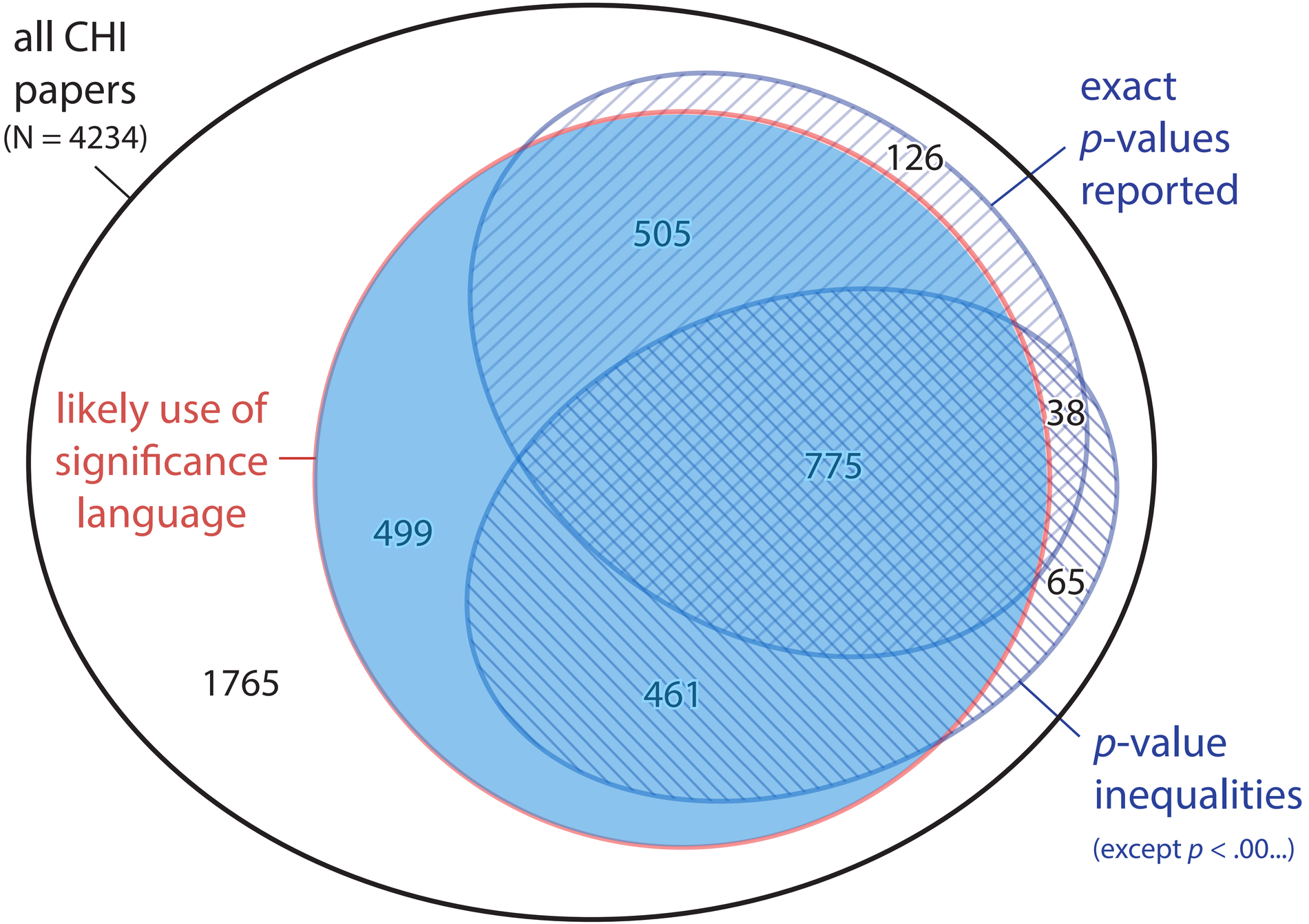
Are we dichotomous in our result interpretation?

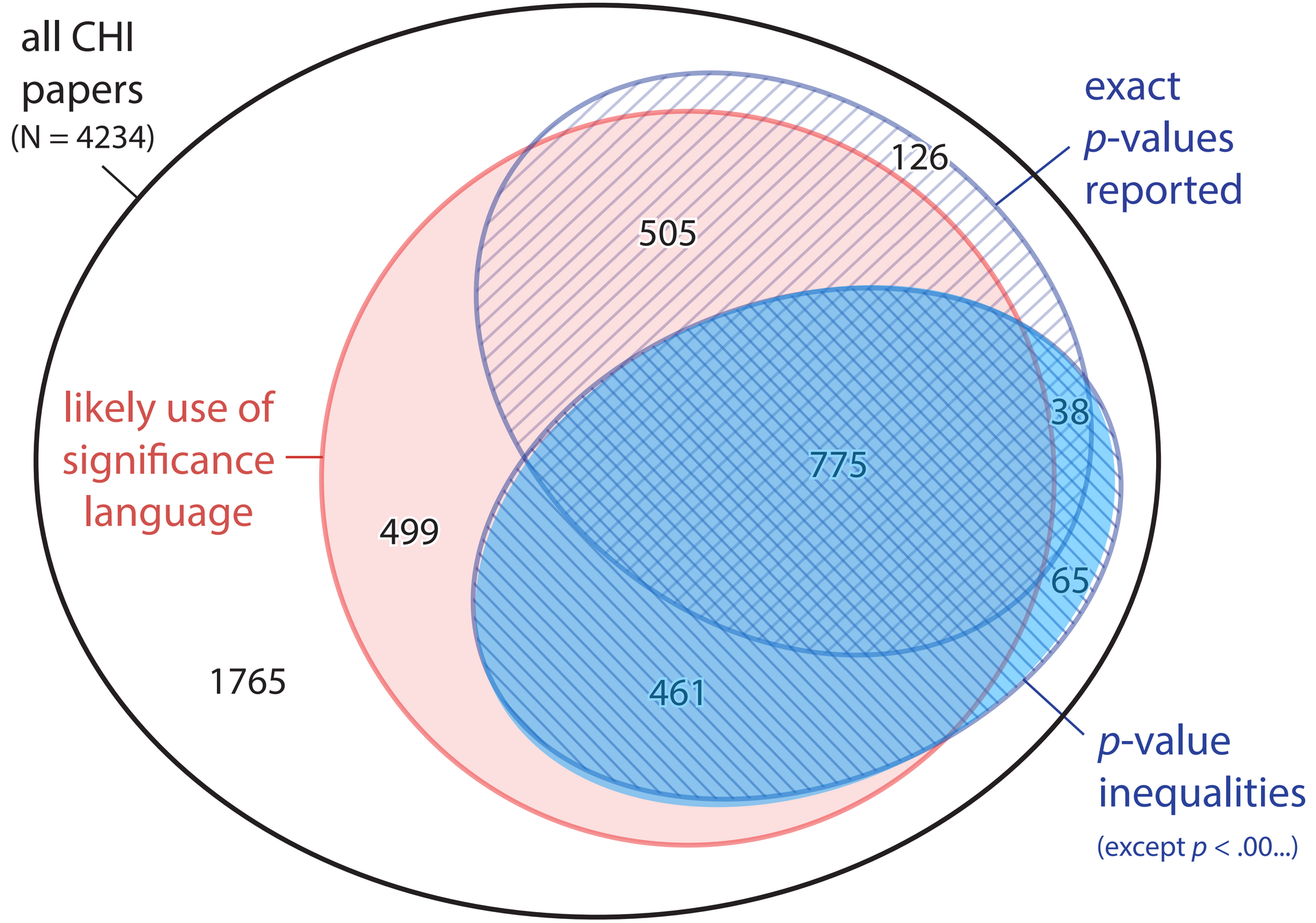

Proportion of CHI papers
100%
75%
50%
25%
0%
2010
2011
2012
2013
2014
2015
2016
2017
2018
likely use of
significance language
use of
"statistically significant"

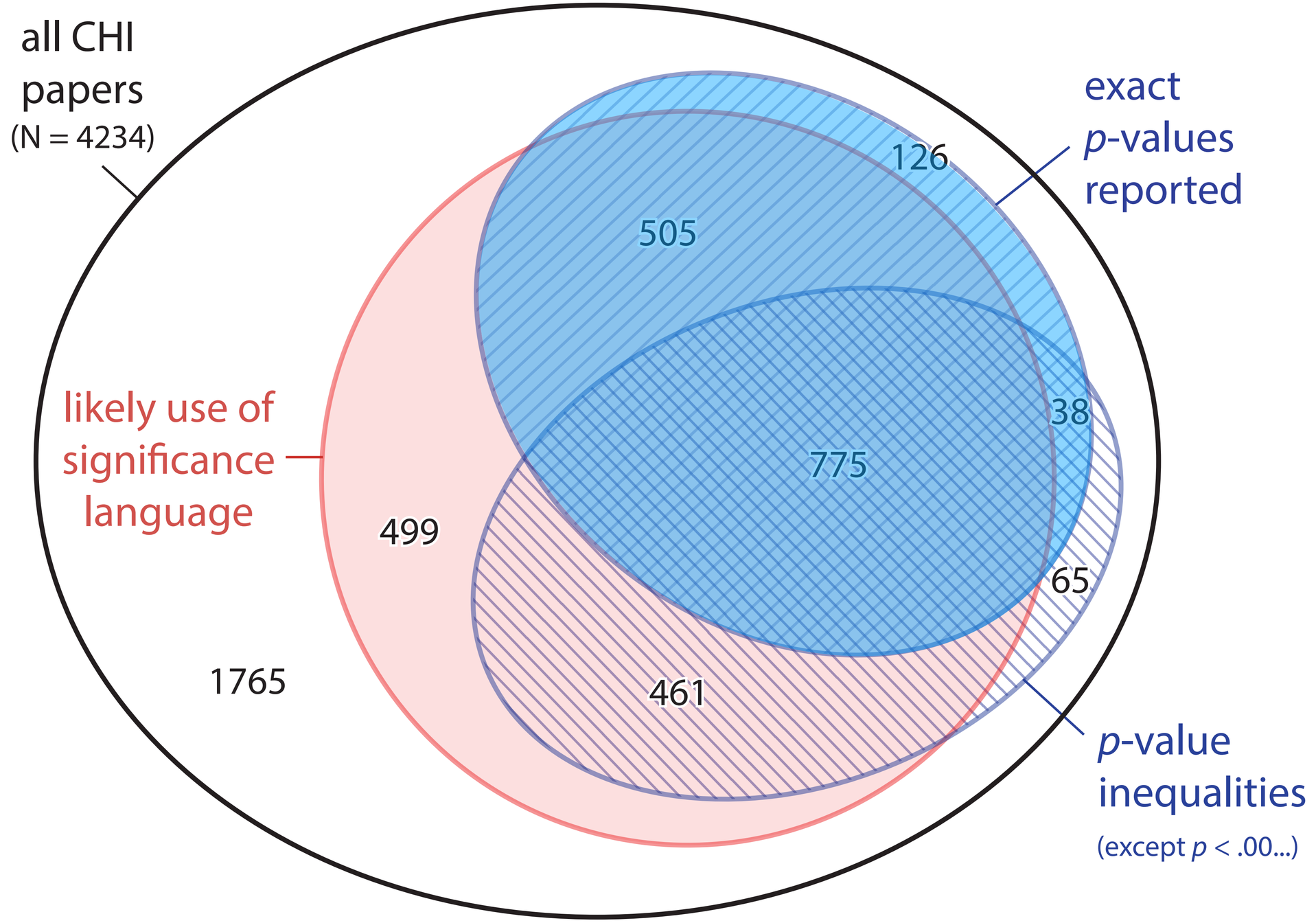
Does the reporting style influence how dichotomous we are in our interpretations?

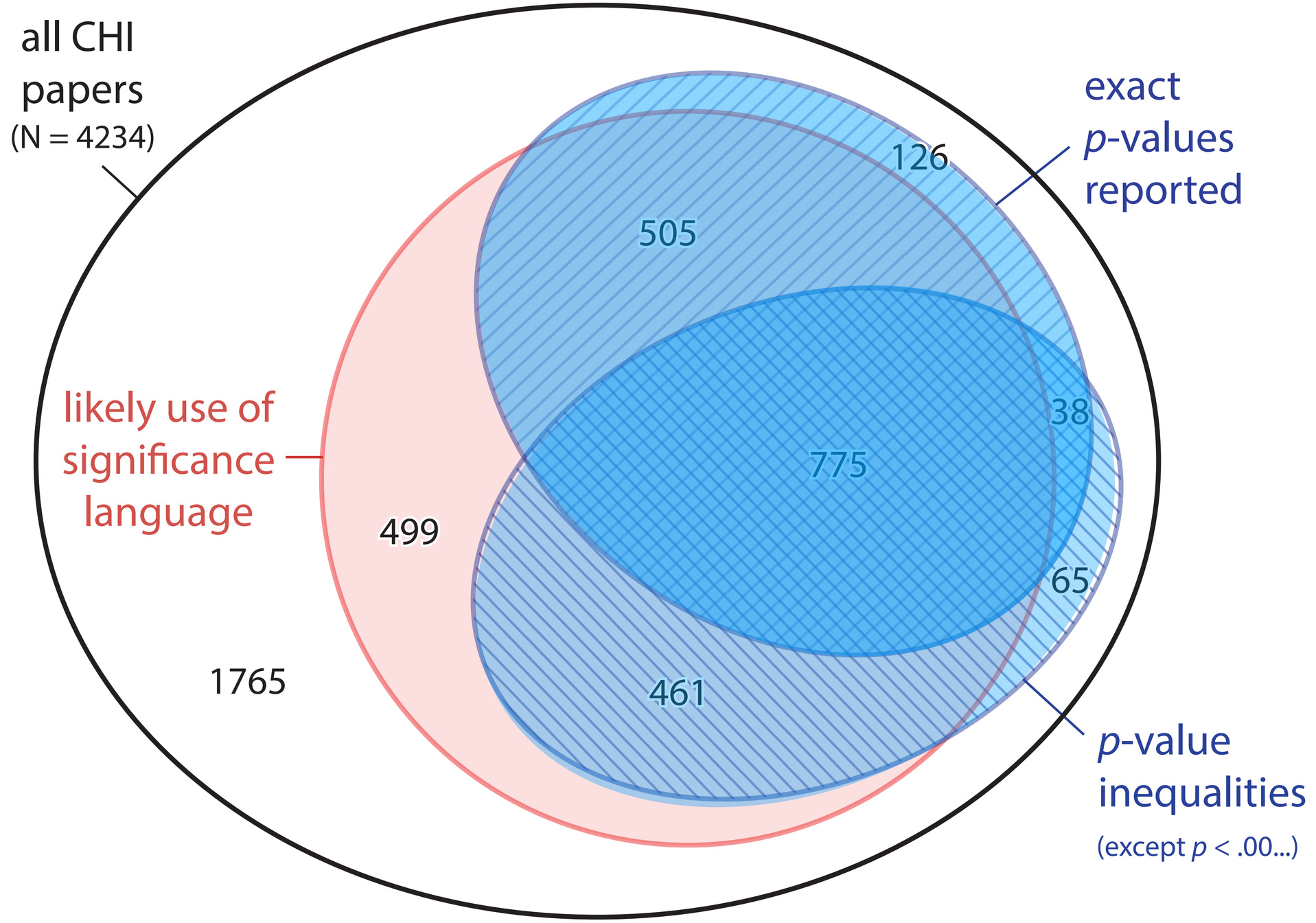

all CHI papers
(N = 4234)
exact p-values reported
likely use of significance language
p-value inequalities
(except p < .00...)
126
505
38
775
499
65
1765
461

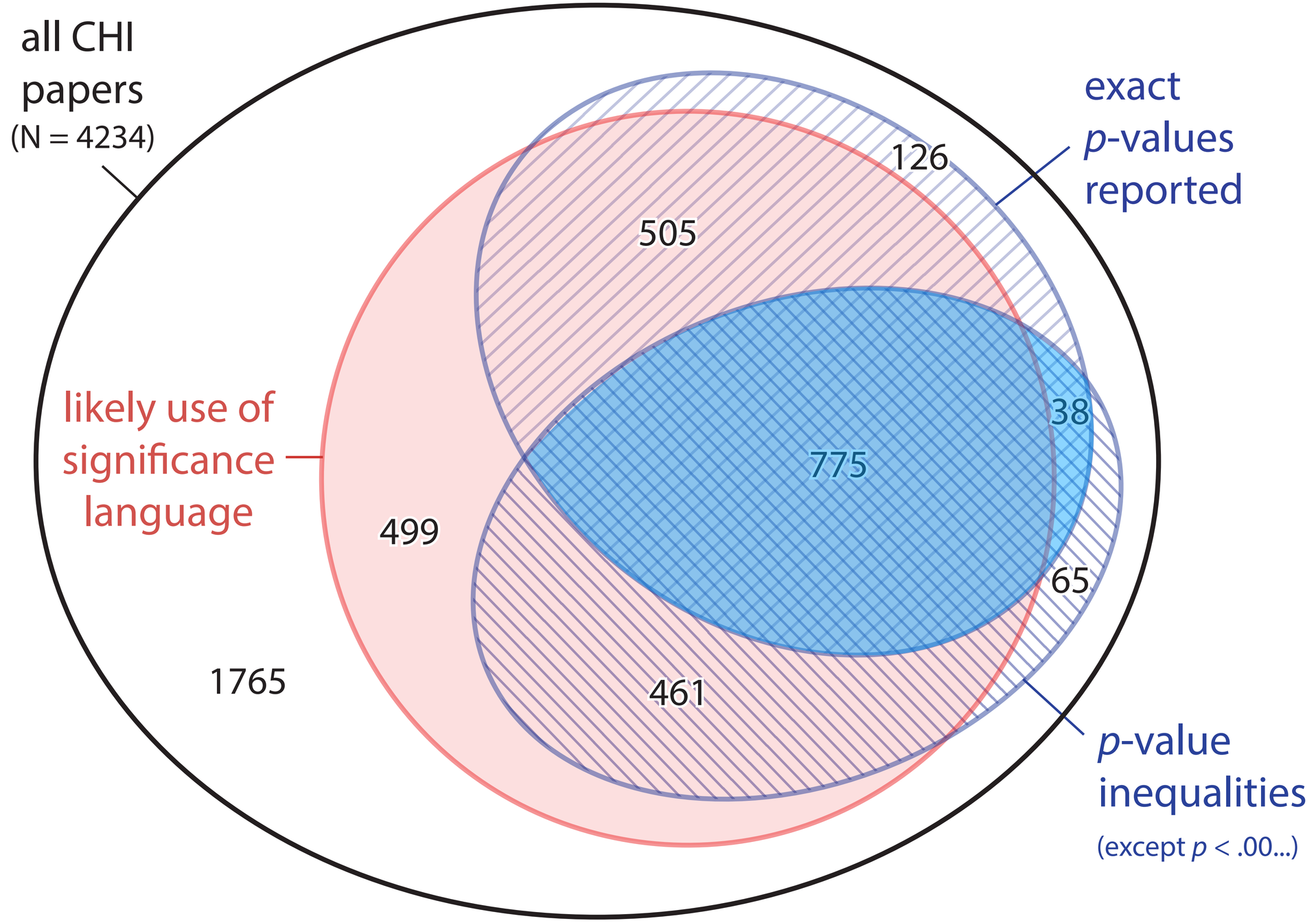

all CHI
papers
(N = 4234)
exact
p-values
reported
126
505
likely use of
significance
language
38
775
499
65
1765
461
p-value
inequalities
(except $p < .00...$)

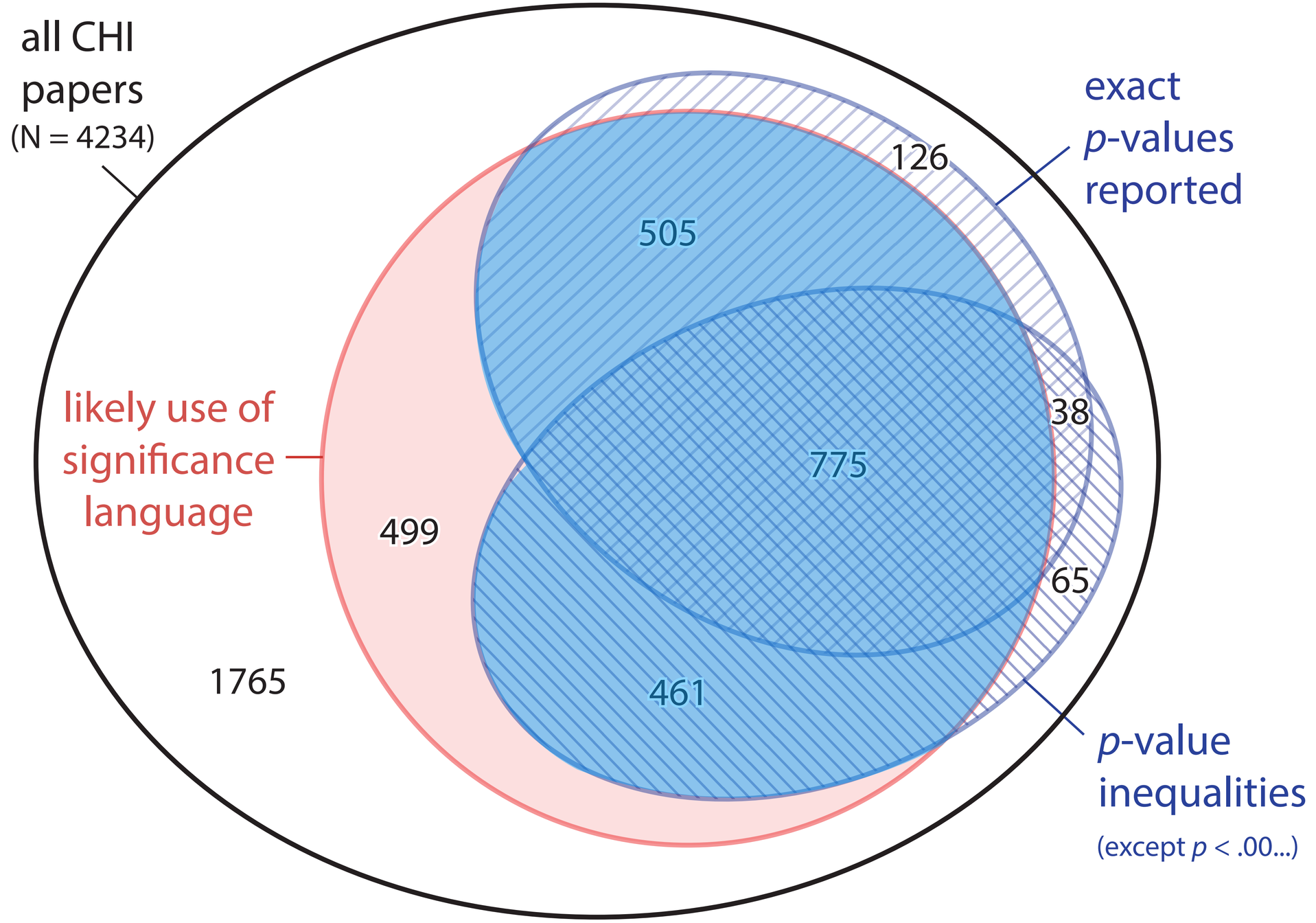

all CHI papers
(N = 4234)
exact p-values reported
likely use of significance language
p-value inequalities
(except p < .00...)
126
505
38
775
499
65
461
1765

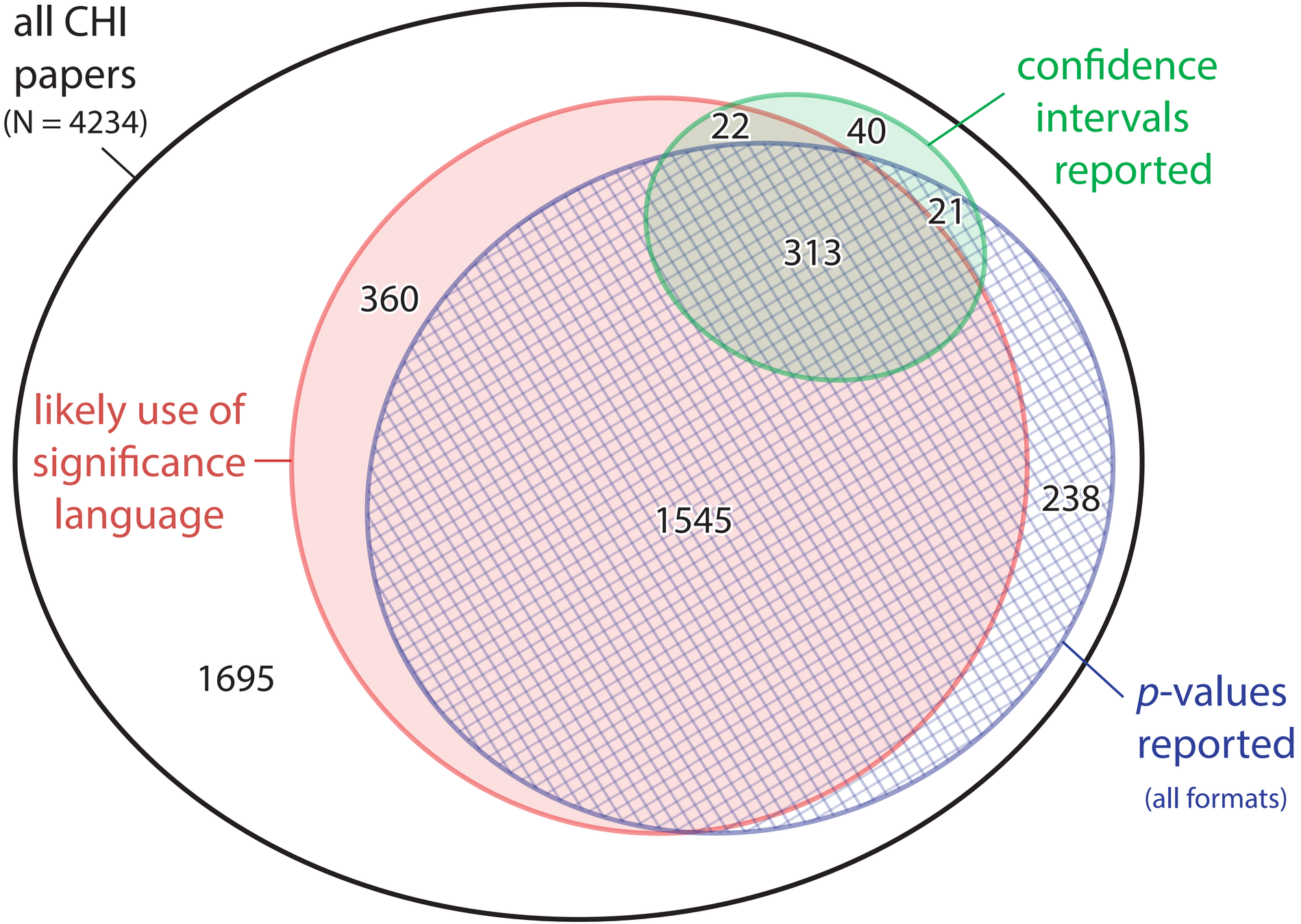

all CHI papers
(N = 4234)
exact p-values reported
126
505
likely use of significance language
499
38
775
65
1765
461
p-value inequalities
(except p < .00...)

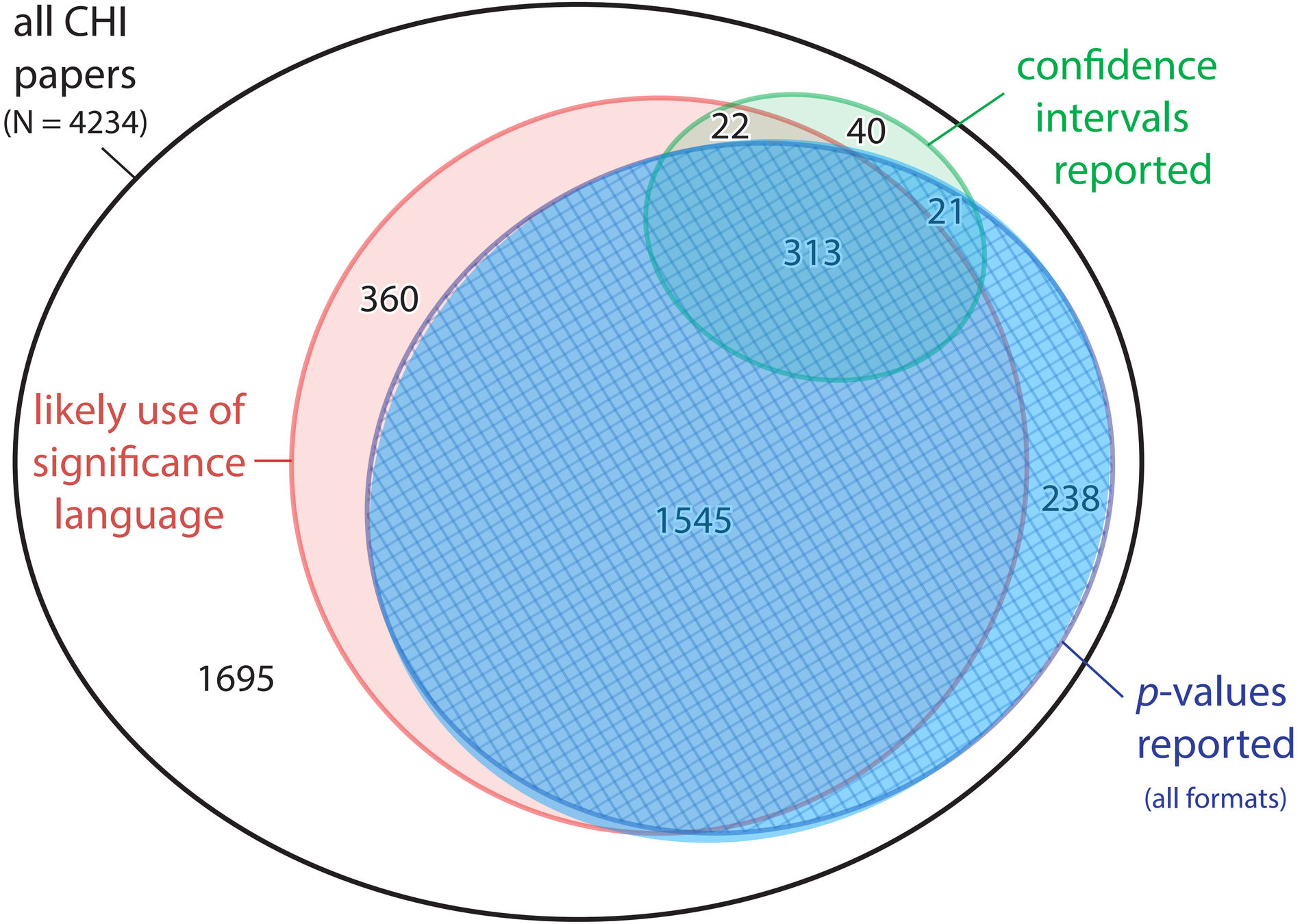

all CHI
papers
(N = 4234)
exact
p-values
reported
126
505
likely use of
significance
language
38
775
499
65
1765
461
p-value
inequalities
(except $p < .00...$)

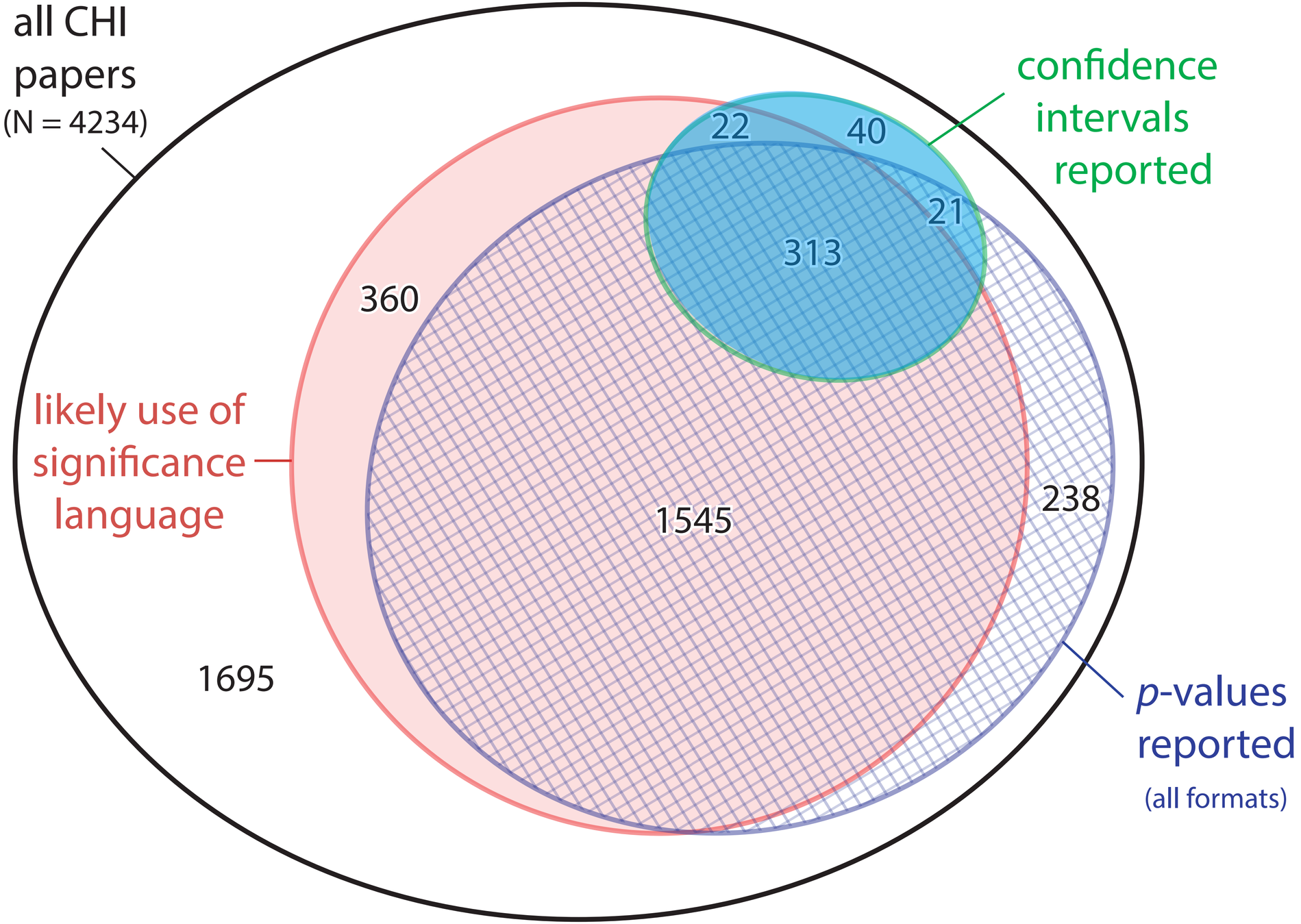

all CHI papers
(N = 4234)
exact p-values reported
126
505
likely use of significance language
499
38
775
65
1765
461
p-value inequalities
(except p < .00...)

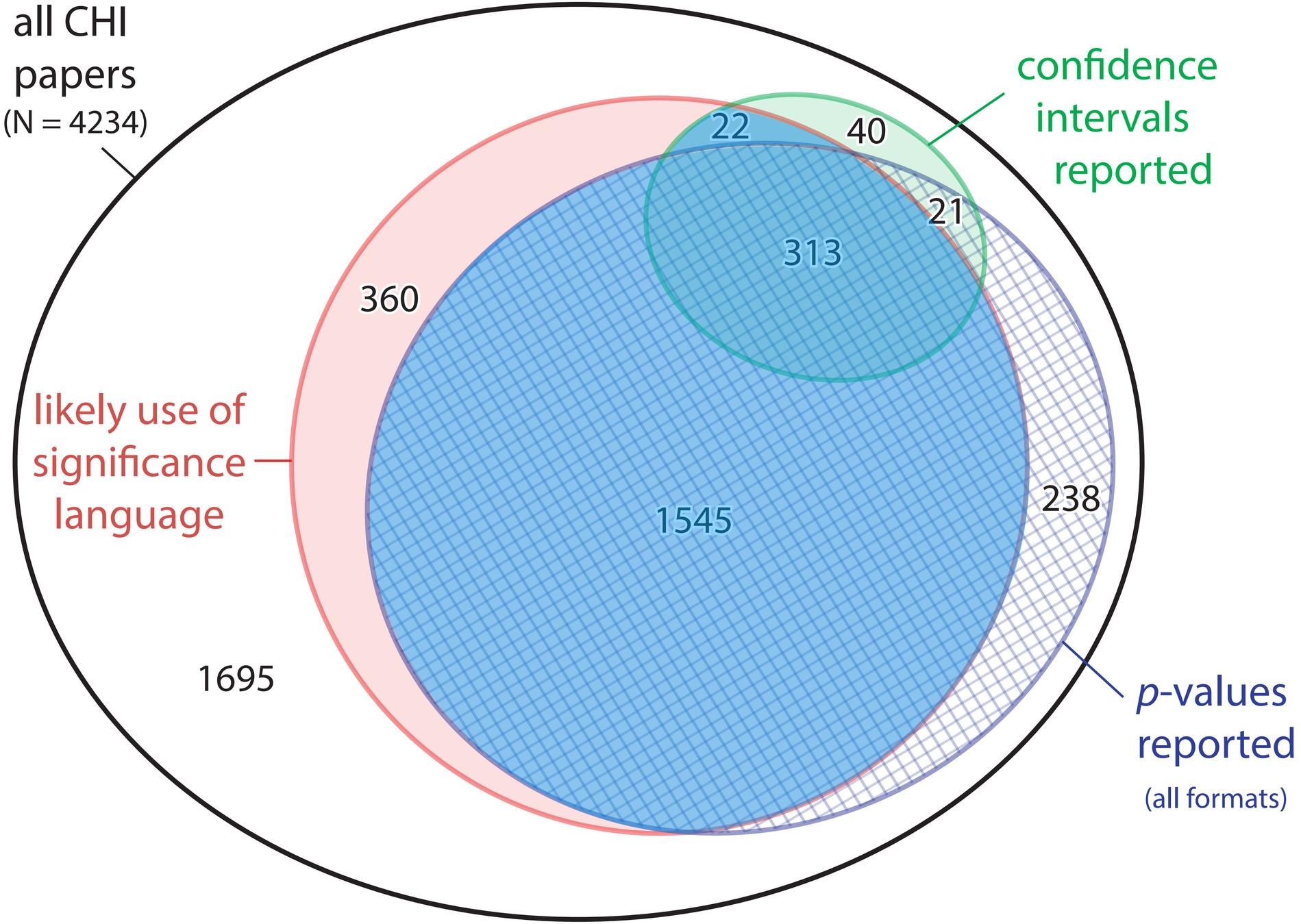

all CHI
papers
(N = 4234)
exact
p-values
reported
126
505
likely use of
significance
language
38
775
499
65
1765
461
p-value
inequalities
(except p < .00...)

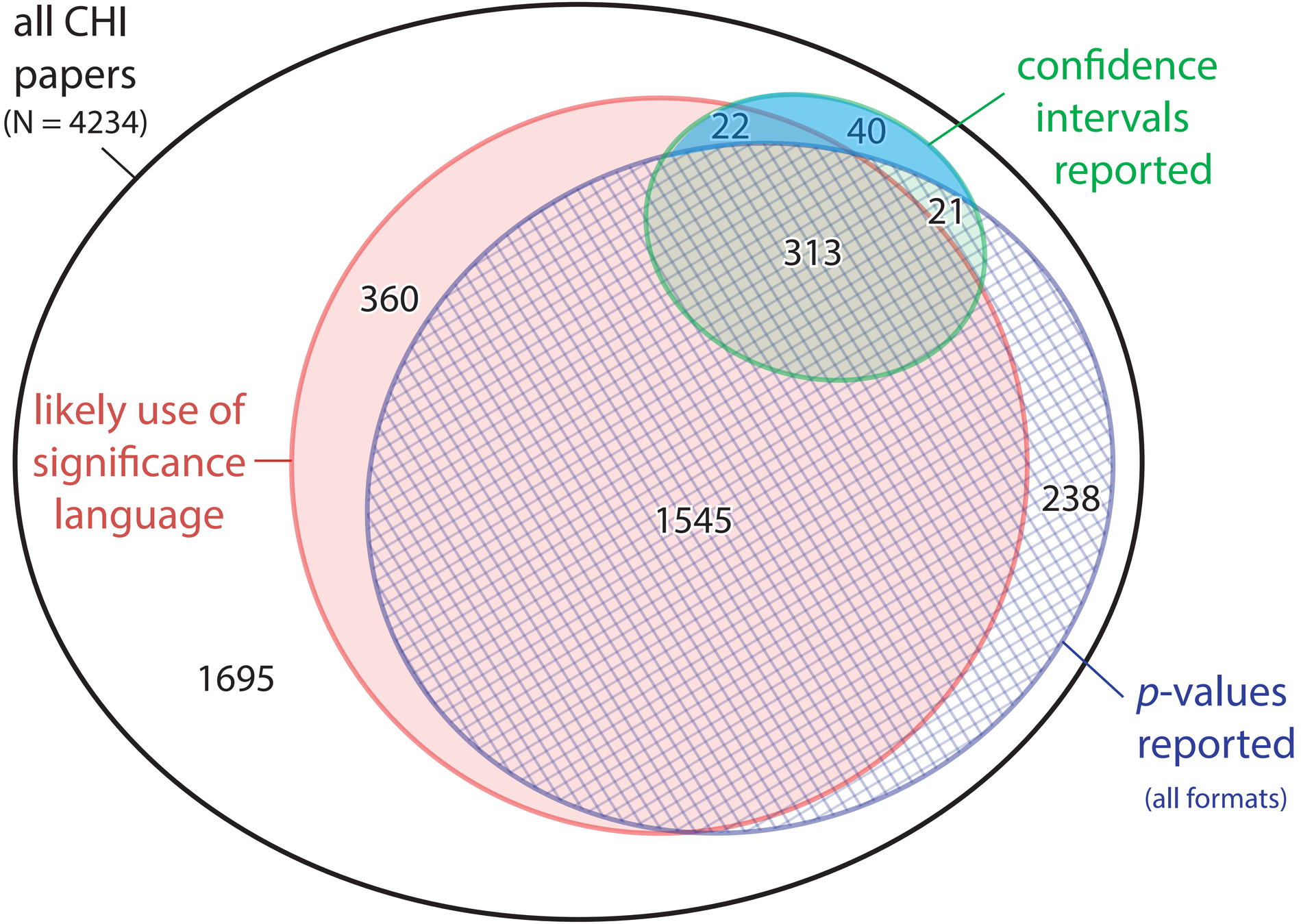

all CHI
papers
(N = 4234)
exact
p-values
reported
126
505
likely use of
significance
language
38
775
499
65
1765
461
p-value
inequalities
(except p < .00...)


all CHI
papers
(N = 4234)
confidence
intervals
reported
22
40
21
313
360
likely use of
significance
language
1545
238
1695
p-values
reported
(all formats)


all CHI
papers
(N = 4234)
confidence
intervals
reported
22
40
21
313
360
likely use of
significance
language
1545
238
1695
p-values
reported
(all formats)


all CHI papers
(N = 4234)
confidence intervals reported
22
40
21
313
360
likely use of significance language
1545
238
1695
p-values reported
(all formats)


all CHI papers
(N = 4234)
confidence intervals reported
likely use of significance language
p-values reported
(all formats)
22
40
21
313
360
1545
238
1695

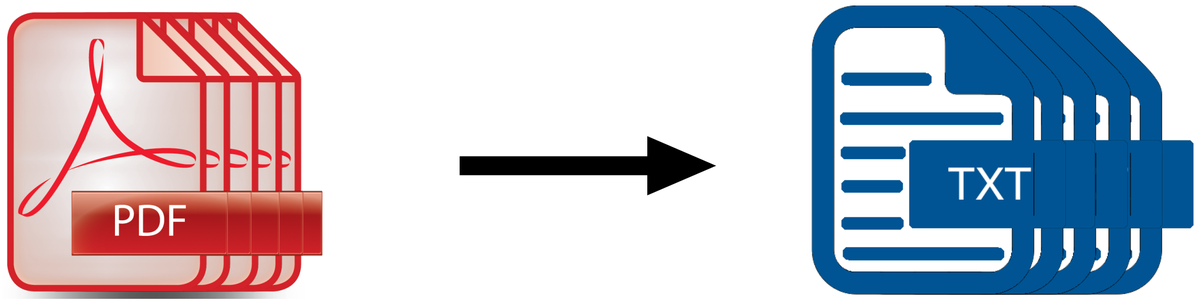

all CHI papers
(N = 4234)
confidence intervals reported
22
40
21
313
360
likely use of significance language
1545
238
1695
p-values reported
(all formats)

The vast majority of papers reporting inferential statistics make dichotomous inferences.

Modest improvement in reporting strategies, but

NHST-based dichotomous inferences have shown no sign of evolution since 2010.

Limitations



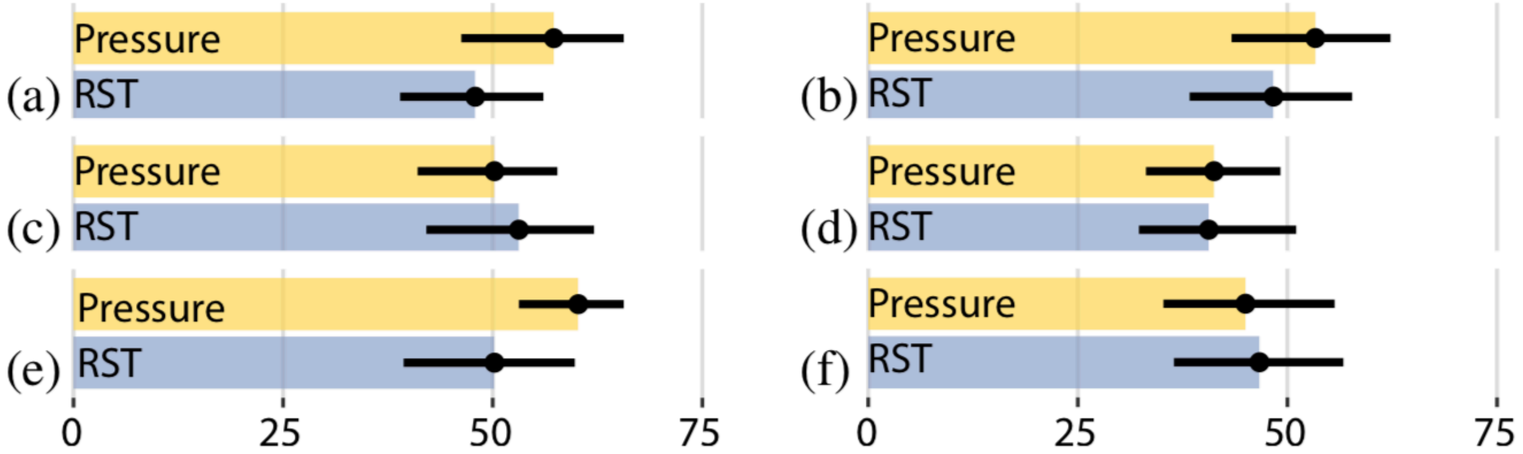
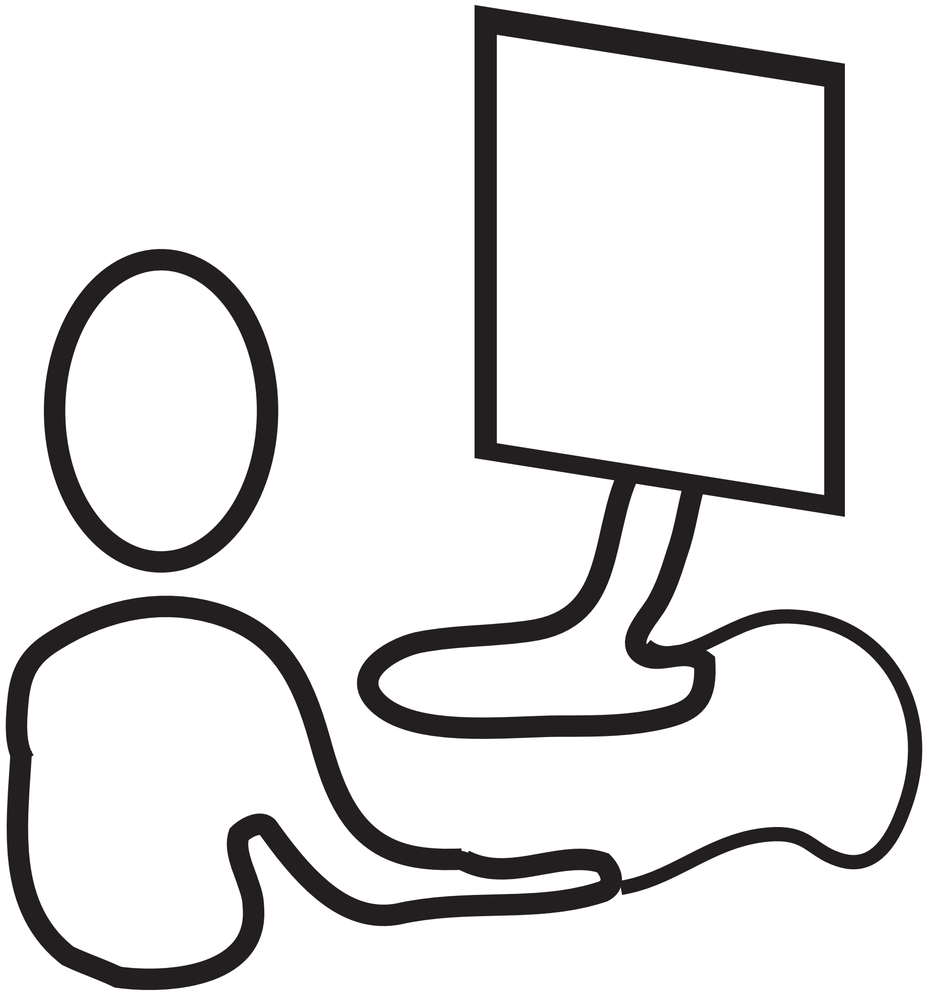


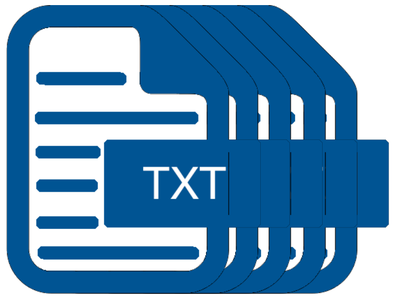
Figure 6: *Workload in TLX units (lower is better) for (a) physical, (b) mental, and (c) temporal demand, (d) performance, (e) effort, (f) frustration. Error bars: 95% bootstrapped CIs.*



Possible
False negatives

List of Likely significance language trigrams

Possible
False positives



List of papers containing likely significance language

Paper #X

“the p-value is significant for speed”

Trigram
“is significant for”

Correctly classified

Paper #Y

“this work is significant for our approach”

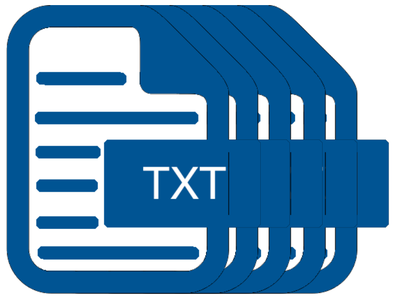
Trigram
“is significant for”

Incorrectly classified

Possible
False negatives

List of Likely significance language trigrams

Possible
False positives



Possible
False negatives

List of papers containing likely significance language

Possible
False positives

“We found that A is faster than B”

“Results indicate that B is the favorite technique”

“There is no difference between C and D”

“Significant”/“Significantly”

“B is more accurate than A”

“Our results show that Z is preferred”

“We conclude that designers should focus on implementing B over A”

Future work

- Improve automated analysis?

- Improve automated analysis?
- Use manual analysis?

tiny.cc/dichotomoussurvey

- Improve automated analysis?
- Use manual analysis?
- Writing guidelines.

Bad Stats: Not what it Seems

Achieving transparent statistical communication in HCI research

Pierre Dragicevic and colleagues

"In the post p<0.05 era, scientific argumentation is not based on whether a p-value is small enough or not. Attention is paid to effect sizes and confidence intervals. Evidence is thought of as being continuous rather than some sort of dichotomy." *Ron Wasserstein, executive director of the American Statistical Association, 2016.*

This web page provides arguments and reading material to explain why it would be beneficial for human-computer interaction and information visualization to move beyond mindless null hypothesis significance testing (NHST), and focus on presenting informative charts with effect sizes and their interval estimates. Our scientific standards can also be greatly improved by planning analyses and sharing experimental material online. At the bottom of this page you will find studies published at CHI and VIS without any p-value, some of which have received best paper awards.

Table of Contents

News:

- Julie Ducasse's PhD thesis Tabletop tangible maps and diagrams for visually impaired users analyzes all of its studies using estimation and reports no single p-value.
- Lonni Besançon's PhD thesis An interaction Continuum for 3D Dataset Visualization analyzes all of its studies using estimation technique and does not report p-values. Additionally, it presents an appendix justifying for the use of estimation techniques instead of the classical dichotomous interpretation.

- **CHI 2017**
  - The study by Dimara, Bezerianos and Dragicevic Narratives in Crowdsourced Evaluation of Visualizations: A Double-Edged Sword? has no p-value and has experimental material online.
  - The study by Besançon, Issartel, Ammi and Isenberg Mouse, Tactile, and Tangible Input for 3D Manipulation makes no use of p-values and uses plots with confidence intervals instead.
  - The study by Besançon, Ammi and Isenberg Pressure-Based Gain Factor Control for Mobile 3D Interaction using Locally-Coupled Devices makes no use of p-values and uses plots with confidence intervals instead. It received a best paper **honorable mention award**.
  - The study by Boy, Pandey, Emerson, Satterthwaite, Nov, and Bertini Showing People Behind Data: Does Anthropomorphizing Visualizations Elicit More Empathy for Human Rights Data? reports its results using confidence intervals.
- **IHM 2017**
  - Emmanuel Dubois and Marcos Serrano published three studies using estimation only at the French-speaking HCI conference IHM 2017. One study co-authored with Perelman, Picard, and Derras received the **best paper award**. The other two studies were co-authored by Raynal, and by Cabric.
- **VIS 2017**
  - The study by Walny, Huron, Perin, Wun, Pusch, and Carpendale Active Reading of Visualizations uses planned analyses, reports all results using estimation and has experimental material online.
  - The study by Dragicevic and Jansen Blinded with Science or Informed by Charts? A Replication Study uses planned analyses, reports all results using estimation and has experimental material online.
  - The study by Perin, Wun, Pusch, and Carpendale Assessing the Graphical Perception of Time and Speed on 2D+Time Trajectories uses planned analyses, reports all results using estimation and has experimental material online.
  - The study by Hullman, Kay, Kim, and Shrestha Imagining Replications: Graphical Prediction & Discrete Visualizations Improve Recall & Estimation of Effect Uncertainty reports all results using Bayesian estimation and has experimental material online.
  - The study by Felix, Bertini, and Franconeri Taking Word Clouds Apart: An Empirical Investigation of the Design Space for Keyword Summaries uses planned analyses and reports all results using estimation.
  - The study by Dimara, Bezerianos and Dragicevic Conceptual and Methodological Issues in Evaluating Multidimensional Visualizations for Decision Support uses planned analyses and reports all results using estimation.
  - The study by Wang, Chu, Bao, Zhu, Deussen, Chen, and Sedlmair EdWordle: Consistency-preserving Word Cloud Editing reports its results using estimation.
  - The study by Valdez, Ziefle, and Sedlmair Priming and Anchoring Effects in Visualization reports most of its results using estimation.
- **SUI 2017**
  - The study by Li, Willett, Sharlin and Costa Sousa Visibility Perception and Dynamic Viewsheds for Topographic Maps and Models reports all of its results using estimation.
- **CHI 2018**
  - The study by Jansen and Hornbæk How Relevant are Incidental Power Poses for HCI? reports all its results using estimation, has experimental material online, and received a **best paper award**.
  - The study by Fernandes, Walls, Munson, Hullman and Kay Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making reports all its results using estimation, uses partly pre-registered analyses, has experimental material online, and received a **honorable mention award**.
  - The study by Feng, Deng, Peck and Harrison The Effects of Adding Search Functionality to Interactive Visualizations on the Web reports all results using estimation and has experimental material online.
- **Expressive 2018**
  - The study by Besançon, Semmo, Biau, Frachet, Pineau, Sariali, Taouachi, Isenberg, and Dragicevic Reducing Affective Responses to Surgical Images through Color

- Improve automated analysis?
- Use manual analysis?
- Writing guidelines.

aviz.fr/badstats

aviz.fr/dichotomous

@lonnibesancon

lonni.besancon@gmail.com