This item is the archived peer-reviewed author-version of:

COVID-19, the Yule-Simpson paradox and research evaluation

COVID-19, the Yule-Simpson paradox and research evaluation

Zhiqi Wang[1] and Ronald Rousseau[2,3]

Abstract

The Yule-Simpson paradox refers to the fact that outcomes of comparisons between groups are reversed when groups are combined. Using Essential Sciences Indicators (ESI), a part of InCites (Clarivate), data for countries, it is shown that although the Yule-Simpson phenomenon in citation analysis and research evaluation is not common, it isn't extremely rare either. The Yule-Simpson paradox is a phenomenon one should be aware of, otherwise one may encounter unforeseen surprises in scientometric studies.

Keywords Yule-Simpson paradox; Relative citations; Scientometric comparisons between countries

Introduction: COVID-19 victims

This work is meant as an illustration of Simpson's paradox, also known as the Yule-Simpson paradox (Yule 1903; Simpson 1951). We use the COVID-19 pandemics as an occasion to show how some basic mathematical observations apply to many aspects of life, in this case victims of the COVID-19 pandemics and scientific contributions of countries as measured by citations per publication.

[*]Ronald Rousseau (Corresponding Author)
ronald.rousseau@uantwerpen.be; ronald.rousseau@kuleuven.be
Zhiqi Wang
zhiqi_wang90@126.com
1 WISE Lab, Institute of Science of Science and S&T management, Dalian University of Technology, Dalian, P.R. China.
2 University of Antwerp, Faculty of Social Sciences, 2020 Antwerpen, Belgium.
3 Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven, 3000 Leuven, Belgium.

On June 22, 2020, R.R.'s local Flemish newspaper, *De Standaard*, mentioned that in any age group men have a higher COVID-19 infection fatality rate (IFR) than women, but in the total population of Belgium women have a higher IFR, see Table 1, as taken from this newspaper article (De Smet 2020). The IFR is the probability that one dies, given that one is infected.

Table 1 Infection fatality rate in different age groups for men and women in Belgium (June 2020)

Age groups	Men (%)	Women (%)
0-24	0.00	0.00
25-44	0.02	0.01
45-64	0.29	0.14
65-74	2.92	1.61
75-84	5.56	3.35
85 and older	13.2	11.07
All ages	1.18	1.31

The fact that in any age group men have a higher infection fatality rate (IFR) than women, but when bringing all age groups together the opposite is the case, seems to be contradictory. This phenomenon is well-known in statistics and is known as Simpson's paradox. Being a quality newspaper, *De Standaard* also mentioned that the data of Table 1 reflect Simpson's paradox (a term not often used in dailies). In this case, the underlying reason is that there are much more older women than men in Belgium. The reporter got his information from an – yet unpublished - article by Flemish colleagues (Molenberghs et al. 2020). Data shown in Table 1 have been updated recently, but do not detract from the fact that in June 2020 the best available data showed the Yule-Simpson paradox.

The Yule-Simpson paradox

Simpson's theoretical example

The Yule-Simpson paradox (Yule 1903; Simpson 1951; Blyth 1972; Gardner 1976) is an expression of a counter-intuitive result that may

occur in statistical aggregations. The paradox refers to the fact that outcomes of comparisons between groups are reversed when groups are combined. Real-world examples have been observed in surgery (Charig et al. 1986), clinical trials (Rücker and Schumacher 2008), ecological studies (Allison and Goldberg 2002; Clark et al. 2011) and, citation analysis (Ramanana-Rahary et al. 2009), among others.

Let us consider the following example shown by Simpson (1951). It appears that the two sets of data separately support a certain hypothesis, but, considered together, support the opposite hypothesis. Simpson provided the following fictitious case related to the outcome of a medical treatment (Table 2).

Table 2 Simpson's survival data

	Male	Female	Total
Untreated	4/7 = 57%	2/5= 40%	6/12 = 50%
Treated	8/13 = 62%	12/27 = 44%	20/40 = 50%

There are 52 cases in total. Among the male population 4/7 ≈ 0.57 of the untreated survived, while 8/13 ≈ 0.62 of the treated ones did. Hence the treatment had a positive effect among males. Among the females, 2/5 = 0.4 of the untreated survived, while 12/27 ≈ 0.44 of the treated ones did. So, also among the female population, the treatment had a positive effect. However if we consider the whole population (bringing males and females together) we see that among the untreated ones 6 survived and 6 died, and among the treated ones 20 survived and 20 died, pointing at no effect from the treatment.

The general framework
The Yule-Simpson paradox occurs in the following situation. Three stochastic variables are involved: X, Y, and Z. In Simpson's example X takes two values: surviving or not surviving; Y also takes two values: being treated or not; and Z represents males or females (these are the ones that are brought together). For the COVID-19 case, X represents

dying or not from COVID-19; Y represents males and females, and Z represents different age groups (here the age groups are brought together).

Now the Yule-Simpson paradox occurs if the following happens (Blyth, 1972): X takes values A and A' (the complement of A); Y takes values B and B' (the complement of B); Z takes values C_1, C_2, C_3, ... (and if there are only two outcomes possible, we denote them by C and C').

For all j= 1,2,3,...: $P(A \mid B$ and $C_j) > P(A \mid B'$ and $C_j)$

and yet: $P(A \mid B) \leq P(A \mid B')$

Here P (. | .) represents a conditional probability. We also say that the Yule-Simpson paradox occurs if:

For all j= 1,2,3,...: $P(A \mid B$ and $C_j) \geq P(A \mid B'$ and $C_j)$

and yet: $P(A \mid B) < P(A \mid B')$

Intuitively one might think that as P(A|B) is an average of the P(A| B and C_j) and similarly for P(A|B') and the P(A|B' and C_j) the paradox is not possible. Yet, the point is that these averages have different weightings (Blyth, 1972). We further note that if Y and Z are independent then the Yule-Simpson paradox is not possible (Blyth, 1972).

The Yule-Simpson paradox and its interpretation can be illustrated graphically using so-called median fractions. As we did that already in our previous article (Ramanana-Rahary et al., 2009), published in this journal, we refer the interested reader to that publication.

A short overview of some historical cases of the Yule-Simpson paradox

As suggested by a reviewer we provide some details on other historical cases of the Yule-Simpson paradox.

A well-known historical example relates to tuberculosis deaths in 1910. Referring to Cohen and Nagel (1934, page 449), Wagner (1982) shows that although the overall mortality rate was lower in New York City than in

Richmond (VA), the opposite held when data were stratified into whites and non-whites.

One of the best-known examples of the Yule-Simpson paradox is a study of possible gender bias among graduate school admissions in 1973 to the University of California, Berkeley. On the whole, male students were more likely than female ones to be admitted. However, when examining the individual departments, it appeared that six out of 85 departments were significantly biased against male applicants, whereas four were significantly biased against female ones. A detailed study of the data by Bickel et al. (1975) revealed that female students tended to apply to more competitive departments with low rates of admission whereas men tended to apply to less competitive departments with high rates of admission. It was concluded that there was no bias from the side of the university, but a selection bias on the part of the applicants.

Julious and Mullee (1994) analyzed data obtained by Charig et. al. (1986) on the efficiency of two treatments to remove kidney stones (open surgery vs. percutaneous nephrolithotomy). The new technique proved successful on the whole, but stratification by the size of the kidney stones led to different conclusions. The confounding factor was that surgeons' choice of treatment was not random but influenced by the size of the stone. This example supported the necessity to use random trials.

We next illustrate this example with the real data in the form of a contingency table (Table 3).

Table 3. Successful removal of kidney stones (Charig et al., 1986)

	Large stones	Small stones	All
Open surgery	192/263 = 73%	81/87= 93%	273/350 = 78%
Percutaneous nephrolithotomy	55/80 = 62%	234/270 = 83%	289/350 = 83%

The stochastic variable X takes the values successful or not; Y takes the values open surgery or percutaneous nephrolithotomy, and Z takes the values large stones or small ones.

Finally, we discuss Yule's original example (Yule 1903, p. 133). This case is related to the study of inheritance and is formulated differently. Yule provides (fictitious) data on a trait that is not hereditary in the male line, and neither in the female line, but occurs with a different probability. Bringing data together in equal proportions suggests inheritance, which is the wrong conclusion.

Table 4. Male line (has or does not have the trait)

	Son has	Son does not have
Father has	25	25
Father does not have	25	25

Table 5. Female line (has or does not have the trait)

	Daughter has	Daughter does not have
Mother has	1	9
Mother does not have	9	81

Table 6. Mixed (sum table)

	Offspring has	Offspring does not have
Parent has	26	34
Parent does not have	34	106

Whether or not the father has the trait, the probability that his son has it is 25/50 = 50%; whether or not the mother has the trait, the probability that her daughter has it is 1/10 = 9/90 = 10%. Yet making these calculations in the sum table yields 26/60 = 43% and 34/140 = 24%. Yule writes that a large but illusory inheritance is created simply by mixing the

two distinct records. He then warns against pooling data about heterogeneous material in general.

Mittal (1991) refers to this form of the paradox as Yule's Association Paradox, while he refers to the case shown by Simpson (1951) as Yule's Reversal Paradox (because the signs in the aggregated table are reversed). Mittal (1991) quotes Nagel as the source for attaching the name of Yule to these two types of paradoxes.

An interpretation related to impact

Direct impact: a fictitious example

The boxes in Table 7, taken from (Ramanana-Rahary et al. 2009), represent direct impact (citations per publication). Research is performed by two countries in two related disciplines. We add a row for 'All countries' (here the two countries). We see that Country 2 is better than Country 1 in Discipline 1 as well as in Discipline 2. Yet adding the results leads to the opposite conclusion.

Table 7 An example of direct impact

Direct impact (fractions)	Discipline 1	Discipline 2	Total: all disciplines
Country 1	60 / 100 = 0.60	1 / 10 = 0.10	61 / 110 = 0.55
Country 2	9 / 10 = 0.90	30 / 100 = 0.30	39 / 110 = 0.35
All countries	69 / 110 = 0.63	31 / 110 = 0.28	100 / 220 = 0.45

Relative impact: an example

The example above also produces an inversion for relative impacts. If, instead of comparing Country 1 and Country 2, we compare each country separately to their aggregate 'All countries', say "the World', we see that in the above example: (Score Country 1) < (Score all countries) in Disciplines 1 and 2, i.e. the relative impact with respect to The World, of Country 1 is inferior to unity, but (Country 1) > (All countries) for All disciplines, i.e. its world relative impact is superior to unity. For Country 2

the opposite holds. Numerical values of relative impacts are given in Table 8.

Table 8 Relative impacts expressed as fractions

Relative impact	Discipline 1	Discipline 2	All disciplines
Country 1	(60/100) / (69/110) = 0.96	(1/10) / (31/110) = 0.35	(61/110) / (100/220) = 1.22
Country 2	(9/10) / (69/110) = 1.43	(30/100) / (31/110) = 1.06	(39/110) / (100/220) = 0.78
All countries	(69/110) / (69/110) = 1.00	(31/110) / (31/110) = 1.00	(100/220) / (100/220) = 1.00

An abstract framework

Let us put this in an abstract framework. The Yule-Simpson paradox occurs if Table 9 is given, together with the requirements that $A/U < C/W$ and $B/V < D/X$ while $(A+B)/(U+V) \geq (C+D)/(W+X)$.

Table 9 General framework for the Yule-Simpson paradox

	Discipline 1	Discipline 2	All disciplines
Player 1	A/U	B/V	$(A+B)/(U+V)$
Player 2	C/W	D/X	$(C+D)/(W+X)$
All players	$(A+C)/(U+W)$	$(B+D)/(V+X)$	$(A+B+C+D)/(U+V+W+X)$

Note that A, B, C, D, U, V, W, and X are given, not just the numerical values of the fractions. From now on, we assume that the reader understands the Yule-Simpson paradox and hence we will simply refer to it as the Yule-Simpson phenomenon.

We recall from (Ramanana-Rahary et al. 2009) two simple mathematical results, using the general term 'player' instead of 'country'.

Proposition 1

The Yule-Simpson phenomenon is present for the pair (Player 1, Player 2) for direct impact, if and only if it is also present for relative impact.

This follows immediately from the fact that if in an equality both sides are multiplied or divided by the same positive number, the inequality stays invariant.

Proposition 2

The Yule-Simpson phenomenon for the pair (Player1, Player2) is present if and only if it is present for the pair (Player1, Both Players).

Real-world citation examples

The Essential Science Indicators

The examples we will show are retrieved from the Essential Science Indicators (ESI). Data from the Science Citation Index-Expanded (SCIE) and the Social Sciences Citation Index (SSCI) in the Web of Science (WoS) Core Collection are subdivided into 22 broad fields based on publication and citation performance (Essential Science Indicators 2020). These 22 broad fields are shown in the appendix (Table 14). Data, only articles and reviews, cover a rolling 10-year period and include bimonthly updates. For our investigation, it is important to recall that articles are classified according to the journal in which they are published and that each journal is assigned to only one field. Multidisciplinary journals, however, are an exception to this rule: here a reclassification is performed at the paper level, based on an analysis of the cited references. Data were collected in September 2020. We restricted data to countries that have at least 500 publications (over a 10-year period), except for a few cases where we compared a country with all the other countries in the database, for which the Yule-Simpson phenomenon occurs rarely.

The role of a discipline (as in Table 9) is played by one ESI field. We do not intend to be complete and to combine each ESI field with each other ESI field (it makes little sense to combine, e.g., Chemistry with Social Sciences, General). We just provide some examples in fields for which it may be acceptable to combine them.

Differences in impact (cites per publication) are often rather small so that one may say that they are not statistically significant. Yet, we do not

step into the statistical morass of significance testing (Schneider 2015) and just stick to rankings.

Examples where the Yule-Simpson phenomenon occurs are shown in Tables 10-14.

Table 10. Mathematics - Physics

Pair	Fields	Mathematics	Physics	Union of the two fields
	Countries / regions			
1	**India**	57,531/14,412 = 3.99	629,484/62,347 = 10.10	687,015/76,759 = 8.95
	China	499,826/98,963 = 5.05	2,776,267/268,479 = 10.34	3,276,093/367,442 = 8.92
2	Spain	94,778/18,510 = 5.12	788,581/40,532 = 19.46	883,359/59,042 = 14.96
	Canada	93,442/17,060 = 5.48	618,910/31,724 = 19.51	712,352/48,784 = 14.60
3	**Czech Rep.**	21,983/5,265 = 4.18	232,167/13,950 = 16.64	254,150/19,215 = 13.23
	South Africa	15,826/3,457 = 4.58	116,993/6,782 = 17.25	132,819/10,239 = 12.97
4	Thailand	7,760/2,062 = 3.76	54,882/3,804 = 14.43	62,642/5,866 = 10.68
	Romania	37,077/7,548 = 4.91	144,592/9,611 = 15.04	181,669/17,159 = 10.59
5	**Poland**	45,347/11,818 = 3.84	388,627/29,717 = 13.08	433,974/41,535 = 10.45
	Turkey	50,640/11,103 = 4.56	220,212/16,588 = 13.28	433,974/41,535 = 9.78

Table 11. Molecular Biology & Genetics – Neuroscience & Behavior

Pair	Fields	Molecular Biology & Genetics	Neuroscience & Behavior	Union of the two fields
	Countries / regions			
1	**Singapore**	234,333/5,471 = 42.83	61,580/3,041 = 20.25	295,913/8,512 = 34.76
	Finland	216,039/4,654 = 46.42	124,023/5,153 = 24.07	340,062/9,807 = 34.68
2	Northern Ireland	41,464/769 = 53.92	15,953/636 = 25.08	57,417/1,405 = 40.87
	Wales	96,910/1,745 = 55.54	68,065/2,512 = 27.10	164,975/4,257 = 38.75

| 3 | Slovenia | 25,336/757 = 33.47 | 13,042/658 = 19.82 | 38,378/1,415 = 27.12 |
| | Australia | 596,389/17,538 = 34.01 | 516,531/24,153 = 21.39 | 1,112,920/41,691 = 26.69 |

Table 12. Computer Science – Mathematics

Pair	Fields	Computer Science	Mathematics	Union of the two fields
	Countries / regions			
1	The Netherlands	79,050/7,140 = 11.07	25,789/4,508 = 5.72	104,839/11,648 = 9.00
	England	271,827/24,125 = 11.27	130,313/20,992 = 6.21	402,140/45,117 = 8.91
2	Wales	15,328/1,382 = 11.09	5,185/864 = 6.00	20,513/2,246 = 9.13
	Finland	45,640/4,087 = 11.17	17,583/2,880 = 6.11	63,223/6,967 = 9.07
3	Malaysia	40,311/4,275 = 9.43	8,605/1,782 = 4.83	48,916/6,057 = 8.08
	Greece	52,053/5,357 = 9.72	18,302/3,390 = 5.40	70,355/8,747 = 8.04
4	Ireland	24,065/2,540 = 9.47	8,799/1,729 = 5.09	32,864/4,269 = 7.70
	Belgium	46,543/4,853 = 9.59	26,759/4,733 = 5.65	73,302/9,586 = 7.65

Table 13. Clinical Medicine – Molecular Biology & Genetics

Pair	Fields	Clinical Medicine	Molecular Biology & Genetics	Union of the two fields
	Countries / regions			
1	USA	17,699,989/926,526 = 19.10	6,722,408/184,228 = 36.49	24,422,397/1,110,754 = 21.99
	Greece	575,770/29,343 = 19.62	98,950/2,562 = 38.62	674,720/31,905 = 21.15
2	Singapore	389,425/19,658 = 19.81	234,333/5,471 = 42.83	623,758/25,129 = 24.82
	Ireland	404,061/19,041 = 21.22	137,677/3,007 = 45.79	541,738/22,048 = 24.57
3	Israel	577,269/30,091 = 19.18	248,119/6,611 = 37.53	825,388/36,702 = 22.49
	Greece	575,770/29,343 = 19.62	98,950/2,562 = 38.62	674,720/31,905 = 21.15
4	Germany	4,028,859/211,381 = 19.06	1,474,034/44,261 = 33.30	5,502,893/255,642 = 21.53
	Greece	575,770/29,34	98,950/2,562 =	674,720/31,905

| | | 3 = 19.62 | 38.62 | = 21.15 |

Table 14. Clinical Medicine – Biology & Biochemistry

Pair	Fields	Clinical Medicine	Biology & Biochemistry	Union of the two fields
	Countries / regions			
1	Colombia	137,736/6,459 = 21.32	17,711/1,583 = 11.19	155,447/8,042=19.33
	Argentina	238,754/10,681 = 22.35	72,420/5,854 = 12.37	311,174/16,535 = 18.82
2	Philippines	91,065/2,003 = 45.46	4,759/530 = 8.98	95,824/2,533 = 37.83
	Ukraine	79,517/1,562 = 50,91	11,264/995 = 11.32	90,781/2,557 = 35.50
3	Vietnam	73,058/2,200 = 33.21	10,451/1,039 = 10.06	83,509/3,239 = 25.78
	Costa Rica	25,090/673 = 37.28	6,021/569 = 10.58	31,111/1,242 = 25.05
4	Qatar	76,059/3,670 = 20.72	7,770/630 = 12.33	83,829/4,300 = 19.50
	Argentina	238,754/10,681 = 22.35	72,420/5,854 = 12.37	311,174/16,535 = 18.82

Although a rather rare event, we also found examples of the Yule-Simpson phenomenon between a country and all other countries in the database, see Table 15. Here, we included cases with less than 500 publications.

Table 15 Comparisons with the World

Pair	Fields	Molecular Biology & Genetics	Neuroscience & Behavior	Union of the two fields
1	Vietnam	12,590/536 = 23.49	2,896/167 = 17.34	15,486/703 = 22.03
	All others	12,098,390 / 500,861= 24.16	10,034,241/539,131 = 18.61	22,132,631/1,039,992 = 21.28
2	Bangladesh	10,792/460 = 23.46	2,674/165 = 16.21	13,466/625 = 21.55
	All others	12,100,188/500,937 = 24.16	10,034,463/539,133 = 18.61	22,134,651/104,070 = 21.28
	Fields	Computer Science	Mathematics	Union of the two fields

3	Taiwan	121,248/14,414 = 8.41	33,194/7,089 = 4.68	154,442/21,503 = 7.18
	All others	3,362,268/394,445 = 8.52	2,162,699/451,667 = 4.79	5,524,967/846,112 = 6.53
	Fields	Social Sciences, General	Economics & Business	Union of the two fields
4	Ethiopia	18,742/2,389 = 7.85	1,915/193 = 9.92	20,657/2,582 = 8,00
	All others	7,982,499/1,021,260 = 7.82	2,920,014/300,898 = 9.70	10,902,513/1,322,158 = 8.25

Remarks

1. We found several more examples involving countries and fields with fewer publications.

2. The cases shown in this contribution are just examples of a phenomenon that might not be well-known to all colleagues. We did not check the 'correctness' of the data in the used database.

3. Countries that are compared have relative citations that do not differ much, although their absolute numbers of publications and citations may differ considerably. As countries are rarely compared in this way, this leads to unexpected 'relatives'. So we see India and China, the Netherlands and England, Qatar and Argentina and the USA and Greece, to name a few.

4. The ESI categories are disjoint and hence it makes sense to add publications and citations. A similar exercise is not directly possible with WoS categories or SCImago categories.

5. Incites uses whole counting and hence when two countries are compared a part of their data overlap. The Yule-Simpson phenomenon between two countries might or might not occur if fractional counting were used. Moreover, assume that one removes all co-authored articles between two countries then again the Yule-Simpson phenomenon may or may not occur. Indeed, inequalities may reverse when removing joint publications and their citations. Let $A/U < C/W$ as in $400/300 < 500/350$. If now these countries have 200 publications with 100

citations in common and these are removed then we have 300/100 > 400/150 with reversed inequality.

Conclusions

Although the Yule-Simpson phenomenon in citation analysis is not common, it isn't extremely rare either. This is shown in this contribution. It is a phenomenon one should be aware of, otherwise one may encounter unforeseen surprises. Assume, for instance, that it is the scientific aim of a country to do better, citation-wise, than world average in the two related fields F_1 and F_2. Then this aim may be reached for the union of the two fields, but for none of the fields separately. Such a possibility is just a mathematical fact. The COVID-19 example and the historical examples illustrated that the Yule-Simpson phenomenon may occur in any aspect of life.

From the historical examples, we learned that one can make a distinction between two cases. Sometimes, such as in the Berkeley students case and for the kidney stone case, there is a clear (human) selection procedure at work and it makes no sense to aggregate data. Sometimes, as in the COVID-19 example, there is a natural stratification (age groups), but again it is not important at all to collect information on the aggregated data. So, in general, we think that it is not a good idea to aggregate data as it leads to a clear loss of information.

In the citation analysis presented here, we artificially aggregated fields, yet these fields themselves are aggregates, and we did not try to find the relation, e.g., between mathematics and its subfields (algebra, geometry, topology, analysis, etc.). So for citation analysis, the answer to the question "Should one aggregate or not?" depends on the aim of the investigation.

As an aside we showed that in terms of relative citations, i.e. citations per publication, large, well-known countries such as England and the USA may, in some fields, become comparable with smaller ones such as the Netherlands and Greece.

Appendix

Table 16. The 22 broad research fields used in the Essential Science Indicators (ESI)

No	Research Fields	Web of Science Documents	Cites	Cites/ Paper
1	Clinical Medicine	2,942,586	39,526,641	13.43
2	Chemistry	1,826,753	29,215,059	15.99
3	Materials Science	985,562	15,895,770	16.13
4	Engineering	1,491,145	14,079,684	9.44
5	Biology & Biochemistry	773,798	13,633,446	17.62
6	Physics	1,114,358	13,258,475	11.90
7	Molecular Biology & Genetics	501,397	12,110,980	24.15
8	Neuroscience & Behavior	539,298	10,037,137	18.61
9	Social Sciences, General	1,023,649	8,001,241	7.82
10	Environment/Ecology	583,684	7,971,101	13.66
11	Plant & Animal Science	778,100	7,795,841	10.02
12	Geosciences	502,457	6,863,716	13.66
13	Psychiatry/Psychology	449,811	5,785,883	12.86
14	Pharmacology & Toxicology	435,488	5,771,696	13.25
15	Immunology	272,454	5,255,343	19.29
16	Agricultural Sciences	461,003	4,683,342	10.16
17	Microbiology	221,551	3,581,914	16.17
18	Computer Science	408,859	3,483,516	8.52
19	Economics & Business	301,091	2,921,929	9.70
20	Space Science	154,642	2,913,112	18.84
21	Mathematics	458,756	2,195,893	4.79
22	Multidisciplinary	23,406	423,822	18.11

Note: Data are collected in September 2020

References

Allison, V.J., & Goldberg, D.E. (2002). Species-level versus community-level patterns of mycorrhizal dependence on phosphorus: an example of Simpson's paradox. *Functional Ecology*, *16*(3), 346–352.

Bickel, J.P., Hammel, A.E., & O'Connell, J.W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175), 398-404.

Blyth, C.R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association. 67*(338), 364-366.

Charig, C.R., Webb, D.R., Payne, S.R., & Wickham, O.E. (1986). Comparison of treatment of renal calculi by operative surgery, percutaneous nephrolithotomy, and extracorporeal shock wave lithotripsy. *BMJ, 292*(6524), 879–882.

Clark, J.S., Bell, D.M., Hersh, M.H., Kwit, M. C., Moran, E., Salk, C., Stine, A., Valle, D., & Zhu, K. (2011). Individual-scale variation, species-scale differences: inference needed to understand diversity. *Ecology Letters*, *14*(12), 1273–1287.

Cohen, M. R., & Nagel, E. (1934). *An Introduction to Logic and Scientific Methods*. New York: Hartcourt, Brace, and World.

De Smet, D. (2020). Is corona erger dan de griep? (Is corona worse than the flu?) *De Standaard*, 22 June 2020.

Essential Science Indicators. (2020). Essential Science Indicators. *Clarivate Analytics.* Retrieved from https://clarivate.com/webofsciencegroup/solutions/essential-science-indicators/. Accessed September 2020

Gardner, M. (1976). Mathematical games. On the fabric of inductive logic, and some probability paradoxes. *Scientific American, 234*(3), 119-124.

Julious, S. A., & Mullee, M.A. (1994). Confounding and Simpson's paradox. *BMJ*, 309(6967), 1480-1481.

Mittal, Y. (1991). Homogeneity of subpopulations and Simpson's paradox. *Journal of the American Statistical Association*, 86(413), 167-172.

Molenberghs, G., Faes, C., Aerts, J., Theeten, H., Devleesschauwer, B., Bustos Sierra, N., Braeye, T., Renard, F., Herzog, S., Lusyne, P., Van der Heyden, J., Van Oyen, H., Van Damme, P., & Hens, N. (2020). Belgian COVID-19 mortality, excess deaths, number of deaths per

million, and infection fatality rates (8 March − 9 May 2020). *medRxiv preprint doi:* 10.1101/2020.06.20.20136234. Retrieved from: https://doi.org/10.1101/2020.06.20.20136234.

Ramanana-Rahary, S., Zitt, M., & Rousseau, R. (2009). Aggregation properties of relative impact and other classical indicators: convexity issues and the Yule-Simpson paradox. *Scientometrics*, *9*(2), 311-327.

Rücker, G., & Schumacher, M. (2008). Simpson's paradox visualized: The example of the Rosiglitazone meta-analysis. *BMC Medical Research Methodology*, 8, article number 34.

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics*, *102*(1), 411-432.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Series B*, *13*(2), 238–241.

Wagner, C.H. (1982). Simpson's paradox in real life. *The American Statisticia*n, 36(1), 46-48.

Yule G. U. (1903). Notes on the theory of association of attributes of statistics. *Biometrika*, *2*(2), 121-134.