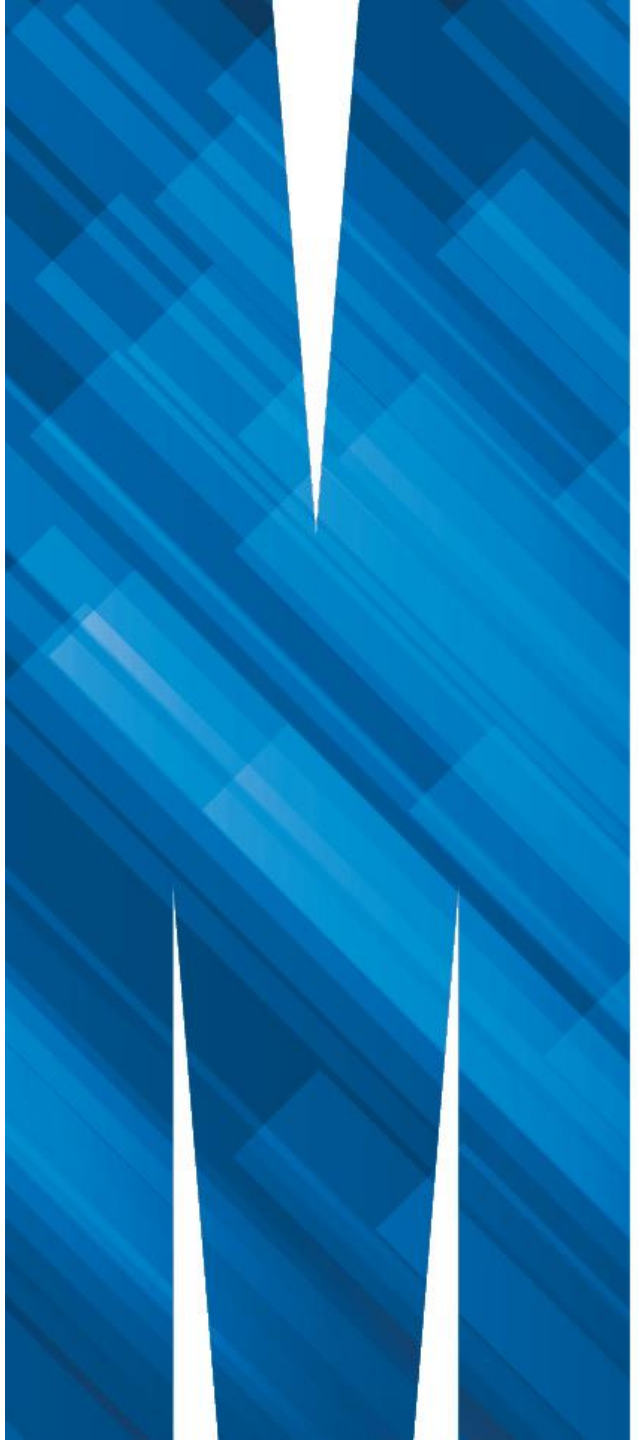
MONASH University

How to assess the scientific integrity of the collected work of one author or group of authors

Jeremy Nielsen, Esmée M Bordewijk, Lyle Gurrin, Jim Thornton, Nicholas J L Brown, Ben W Mol



Why do we need post-publication peer review?

- Carlisle: 14% of trials contain false data, 8% ‘zombie’.
  - 44% false data and 26% zombie for trials providing IPD
- This amounts to an estimated several hundred thousand flawed trials worldwide

- Scientific misconduct distorts the results of evidence syntheses

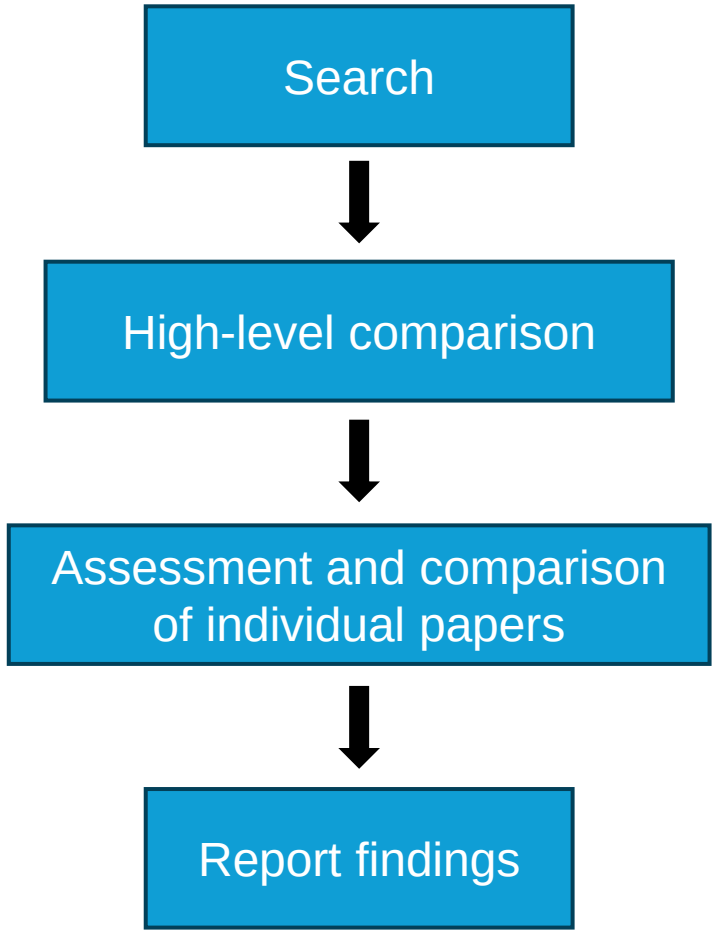
Scientific misconduct is often not a one-time offence

https://retractionwatch.com/the-retraction-watch-leaderboard/

Aim

To develop methods to assess the work of one author or author-group

Systematic integrity assessments



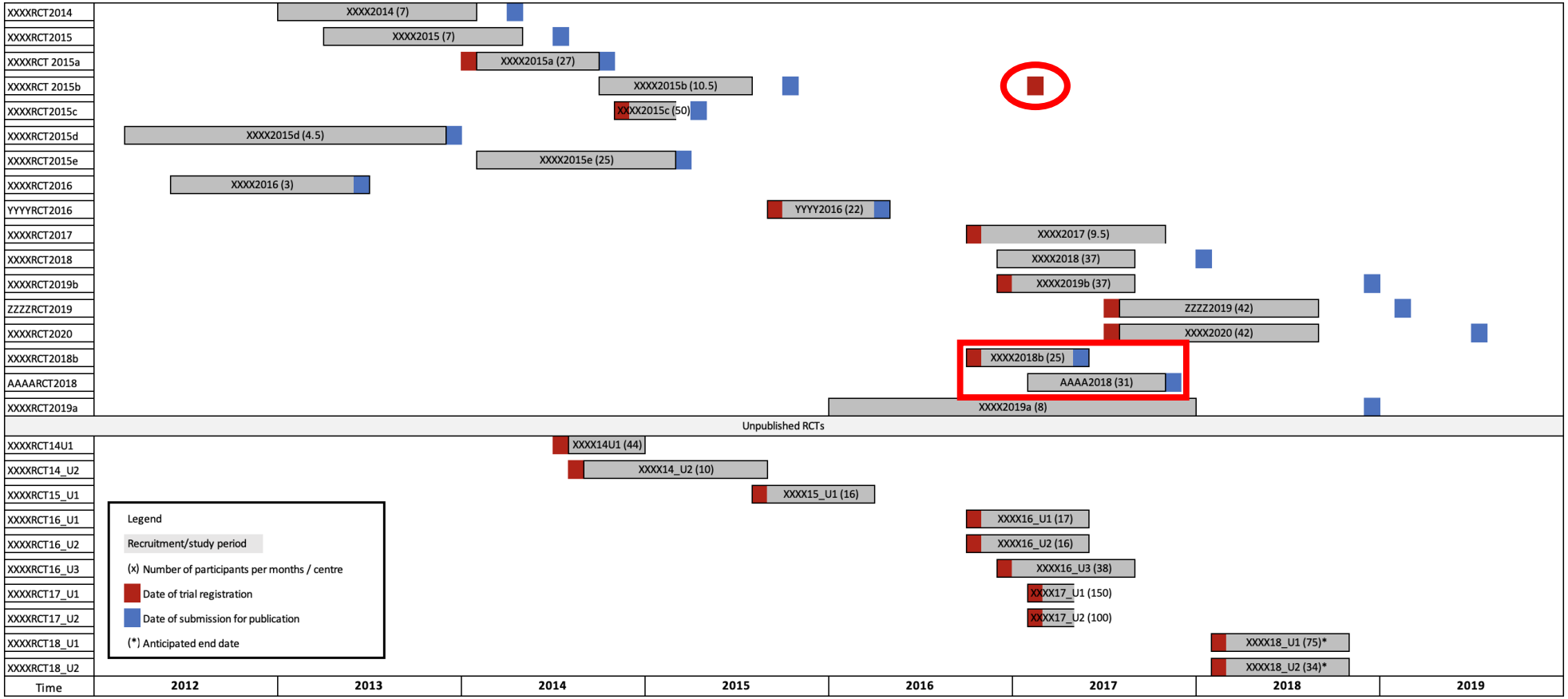
Identifying papers

- PubMed, Google Scholar
- Retracted articles/editorial expressions of concern: RetractionWatch database
- Unpublished studies: clinical trial registries

Identifying papers

Study	Registration	Date registered	Recruitment in trial registry (M-Y)		Recruitment in paper (M-Y)		No. participants planned trial reg.	No. participants randomised	No. participants analysed	No. incl. / month	Date received by journal
			Start	End	Start	End					
XXXX2014	Not registered	NR	NR	NR	Jan 13	Jan 14	NR	40 vs 40	40 vs 40	7	01-04-14
XXXX2015	Not registered	NR	NR	NR	Apr 13	Apr 14	NR	40 vs 40	40 vs 40	7	01-07-14
XXXX2015a	PACTR2014—	31-01-14	05-02-14	31-07-14	Feb 14	Sep 14	200	108 vs 106	100 vs 100	27	23-10-14
XXXX2015b	PACTR2014—	06-11-14	15-11-14	28-02-5	Nov 14	Feb 15	200	118 vs 118	100 vs 100	50	07-04-15
XXXX2015c	Not registered	NR	NR	NR	Mar 12	Nov 13	NR	30/30/30	30/30/30	4.5	25-12-13
XXXX2015d	Not registered	NR	NR	NR	Feb 14	Feb 15	NR	112/116/114	100/100/100	25	20-03-15
XXXX2016	Not registered	NR	NR	NR	Jun 12	Jun 16	NR	54/54/54	50/50/50	3	27-06-16
ZZZZ2016	PACTR2015—	12-09-15	20-09-15	12-04-16	NR	NR	150	72 vs 74	72 vs 74	22	Apr 2016
XXXX2017	PACTR201—	03-10-16	14-10-16	16-10-17	NR	NR	100	57 vs 56	54 vs 52	9.5	NR
XXXX2018	Not registered	NR	NR	NR	Dec 16	Aug 17	330	165 vs 165	152 vs 159	37	05-01-18
XXXX2019	PACTR2017—	01-12-16	14-12-16	18-06-17	Dec 16	Aug 17	300	165 vs 165	152 vs 154	37	05-12-18
YYYY2019	PACTR2017—	24-07-17	01-08-17	01-08-18	Aug 17	Aug 18	480	168/168/168	166/160/164	42	11-02-19
XXXX2020	PACTR2017—	24-07-17	01-08-17	01-08-18	Aug 17	Aug 18	480	165/165/165	164/160/162	42	04-07-19

Study timelines



MONASH University

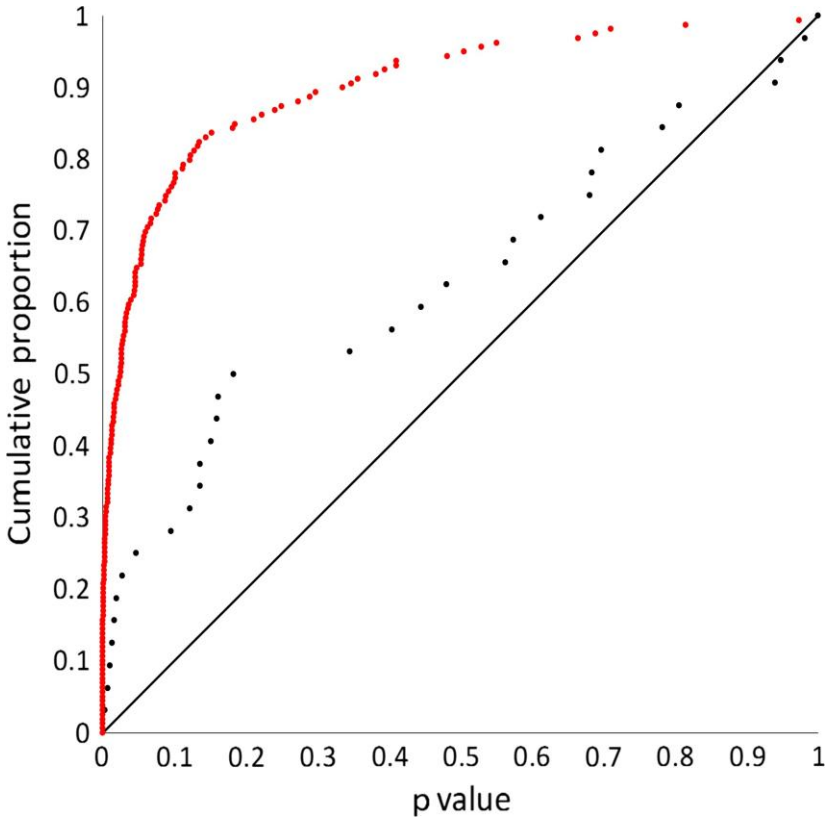
Extracting results

A	B	C	D	E	F	J	K	L	Q	R	S	X	Y	AE	AF	AL	AM	AS	AT
DOI	Table	Item ref	Decimal			Sample 1			Sample 2			Orig *p*	Orig t	Midpoint t		Minimum t		Maximum t	
			DP.M	DP.SD	DP.t	M1	SD1	N1	M2	SD2	N2		(optional)	t	*p*.t	t.Min	*p*.t.Min	t.Max	*p*.t.Max
doi: 10.1016/j	1	Parity	2	2	2	1.76	0.98	93	1.95	1.15	89	0.42	?	-1.20	0.231	-1.133	0.259	-1.271	0.206
doi: 10.1016/j	1	BMI	2	2	2	26.24	1.20	93	26.43	1.64	89	0.63	?	-0.89	0.372	-0.845	0.399	-0.945	0.346
doi: 10.1016/j	1	P level	2	2	2	0.52	0.25	93	11.64	4.53	89	0.00	?	-23.64	0.000	-23.589	0.000	-23.687	0.000
doi: 10.1016/j	1	E2 level	2	2	2	12.84	1.96	93	38.86	3.99	89	0.00	?	-56.21	0.000	-56.099	0.000	-56.313	0.000
doi: 10.1016/j	2	Medio lat Stab index before	2	2	2	2.37	0.53	93	2.42	0.55	89	0.71	?	-0.62	0.533	-0.495	0.621	-0.756	0.450
doi: 10.1016/j	2	Medio lat Stab index after	2	2	2	1.84	0.23	93	2.40	0.56	89	<.0001	?	-8.89	0.000	-8.640	0.000	-9.151	0.000
doi: 10.1016/j	2	Antero-posteriorStab index be	2	2	2	2.38	0.67	93	2.36	0.61	89	0.80	?	0.21	0.834	0.104	0.917	0.318	0.751
doi: 10.1016/j	2	Antero-posteriorStab index af	2	2	2	1.91	0.29	93	2.33	0.61	89	<.0001	?	-5.97	0.000	-5.773	0.000	-6.176	0.000
doi: 10.1016/j	2	Overall Stab index before	2	2	2	2.97	0.50	93	2.95	0.52	89	0.80	?	0.26	0.792	0.131	0.896	0.401	0.689
doi: 10.1016/j	2	Overall Stab index after	2	2	2	2.42	0.29	93	2.95	0.53	89	<.0001	?	-8.42	0.000	-8.166	0.000	-8.674	0.000

This spreadsheet is available on request or online at https://steamtraen.blogspot.com/2021/10/a-catastrophic-failure-of-peer-review.html

Carlisle's method

- For randomised controlled trials
- If randomisation is performed correctly, the expected *p*-value distribution should be uniform
- Traditionally only applied to continuous variables



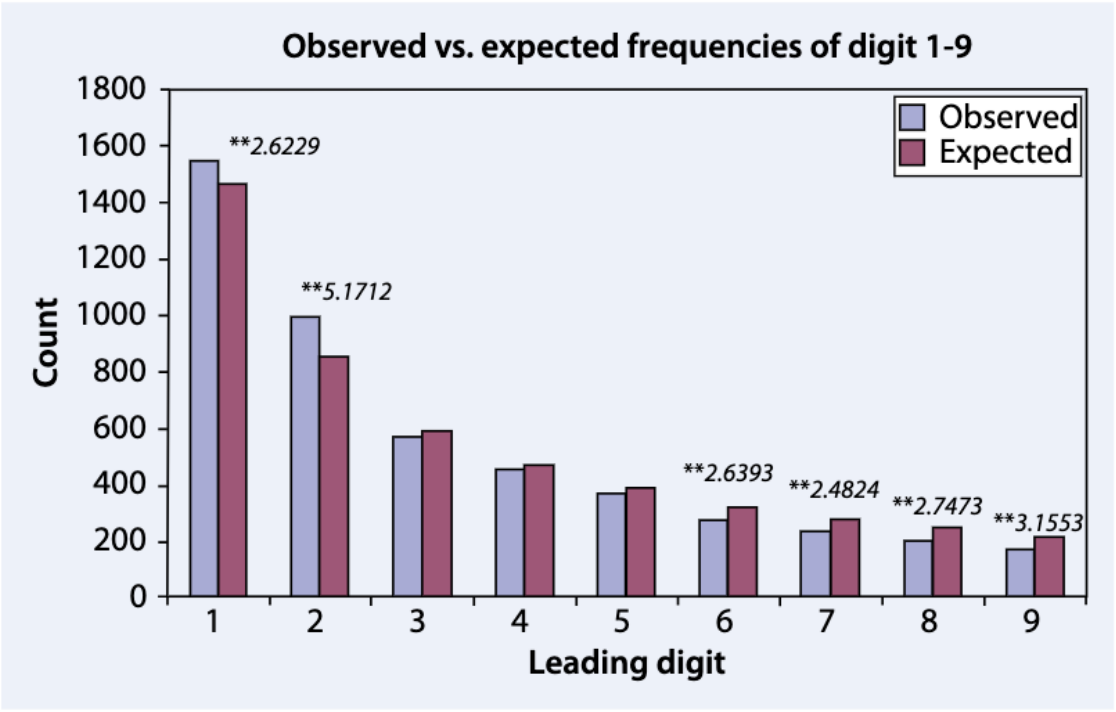
Carlisle & Loadsman. Anaesthesia (2016).

Newcomb-Benford law

Table 1 Frequencies of the first, second, third, fourth, and fifth or greater leading digits according to Benford-Newcomb's law of naturally occurring numbers; results were calculated with formulas 1 and 2, described in the text

Digit	1st (%)	2nd (%)	3rd (%)	4th (%)	5th or higher (%)
0	–	11.97	10.18	10.02	10.00
1	30.10	11.39	10.14	10.01	10.00
2	17.61	10.88	10.10	10.01	10.00
3	12.49	10.43	10.06	10.01	10.00
4	9.69	10.03	10.02	10.00	10.00
5	7.92	9.67	9.98	10.00	10.00
6	6.69	9.34	9.94	9.99	10.00
7	5.80	9.04	9.90	9.99	10.00
8	5.12	8.76	9.86	9.99	10.00
9	4.58	8.50	9.83	9.98	10.00

Hüllemann et al. Anaesthetist (2017).



Hein et al. Anaesthetist (2012).

Comparing results

Table 1. Characteristics of polycystic ovary syndrome patients in the extended clomiphene citrate and gonadotrophin treatment groups.

Parameter	*Clomiphene citrate group*	*Gonadotrophin group (n = 158)*	*95% confidence interval*
No. of cycles (mean no. per patient)	405 (2.53)	397 (2.51)	–
Age (years)	24.1 ± 3.1	26.3 ± 3.0	−0.12 to 0.16
Parity	0.3 ± 0.2	0.3 ± 0.3	−0.30 to 0.05
Height (cm)	160.3 ± 6.2	158.1 ± 5.8	−0.002 to 0.15
Weight (kg)	78.3 ± 6.4	81.1 ± 4.2	−0.26 to 0.45
Clinical presentation			
Oligo/anovulation (%)	136 (85.0)	140 (88.6)	0.36 to 1.3
Hyperandrogenism (%)	76 (47.5)	70 (44.3)	0.63 to 1.99
Polycystic ovaries (%)	111 (69.4)	103 (65.2)	1.05 to 1.44
BMI (kg/m^2)	30.5 ± 3.1	32.5 ± 2.9	−0.02 to 5.4
FSH (IU/ml)	4.1 ± 2.7	5.1 ± 2.1	−0.07 to 2.3
LH (IU/ml)	10.9 ± 1.8[a]	13.1 ± 2.2[a]	−0.07 to 2.5

Values are mean ± SEM unless otherwise stated; BMI = body mass index.
[a] $P = 0.04$. There were no other statistically significant differences.

TABLE 1

Patients' characteristics.

Parameter	Anastrozole group (n = 115)	CC group (n = 101)	Values of χ^2 or t[a]	*P* value	CI
No. of cycles	243	226			
Age (y)	23.8 ± 3.1	25.3 ± 2.9	1.04	.67	−0.12–0.15
Parity	0.3 ± 0.12	0.3 ± 0.16	0.98	.71	−0.30–0.06
Height (cm)	158.3 ± 5.12	155.1 ± 4.20	1.65	.08	−0.002–0.15
Weight (kg)	80.3 ± 5.42	79.1 ± 4.22	0.22	.95	−0.26–0.45
Clinical presentation, n (%)					
Oligoovulation or anovulation	110 (95.6)	92 (91.0)	1.84[b]	1.75	0.36–1.31
Hyperandrogenism	51 (44.3)	42 (41.5)	0.17[b]	.68	0.63–1.99
Polycystic ovaries	98 (85.2)	71 (70.2)	7[b]	.008[c]	1.05–1.41
BMI (kg/m^2)	31.1 ± 2.91	29.1 ± 3.12	1.4	.31	−0.02–5.4
FSH (IU/mL)	6.1 ± 2.92	6.3± 2.22	2.43	.06	−0.07–2.1
LH (IU/mL)	13.2 ± 1.82	12.1 ± 3.11	2.55	.052	−0.06–3.2

[a] Data are *t* values unless otherwise indicated.
[b] Data are χ^2.
[c] Statistically significant difference at $P < .01$.

Badawy. Clomiphene citrate or anastrozole for ovulation induction in women with PCOS. Fertil Steril 2009.

TABLE 2

Outcome in letrozole and anastrozole groups.

	Letrozole group (n = 111)	Anastrozole Group (n = 119)	*t*-test	*P* value
Total number of follicles	5.4 ± 0.4	5.8 ± 0.4	5.21	.01[a]
Number of follicles >14 mm	3.1± 0.3	2.7 ± 0.2	5.33	.004[a]
Number of follicles >18 mm	2.3 ± 0.1	3.1 ± 0.2	8.62	.001[a]
Pretreatment endometrial thickness (mm)	5.5 ± 0.5	5.3 ± 0.6	1.31	.22
Endometrial thickness at hCG (mm)	9.1 ± 0.2	10.2 ± 0.7	4.45	.04[a]
Serum E_2 (pg/mL)	455.1 ± 64.2	484 ± 91.3	2.39	.08
Serum P (ng/mL)	9.2 ± 0.9	10.1 ± 1.2	2.81	.06
Duration of stimulation (days)	11.9 ± 1.3	10.8 ± 2.2	2.30	.21
Pregnancy/cycle	36/295 (12.2%)	42/279 (15.1%)	0.99	.31
Miscarriage/patient	4 (11.1%)	4(9.5%)	0.01	.92

[a] Statistically significant differences as $P < .05$.

Badawy. Letrozole versus anastrozole. Fertil Steril 2008.

TABLE 2

Outcome in letrozole and clomiphene citrate (CC) groups.

	Letrozole group (n = 218)	CC group (n = 220)	*t*	*P* value
Total number of follicles	4.4 ± 0.4	6.8 ± 0.3	4.3	.042[a]
Number of follicles >14 mm	2.1 ± 0.3	3.7 ± 0.5	6.13	.008[a]
Number of follicles >18 mm	2.3 ± 0.1	3.1 ± 0.8	5.03	.03[a]
Pretreatment endometrial thickness (mm)	4.5 ± 0.4	4.3 ± 0.5	1.41	.52
Endometrial thickness at hCG (mm)	8.1 ± 0.2	9.2 ± 0.7	5.44	.021[a]
Serum E_2 (pg/mL)	255.1 ± 64.2	384 ± 91.3	4.12	.022[a]
Serum progesterone (ng/mL)	7.1 ± 0.9	11.1 ± 1.2	6.33	.024[a]
Duration of stimulation (days)	12.1 ± 1.38	8 ± 2.9	4.91	.036[a]
Pregnancy/cycle	82/540 (15.1%)	94/523 (17.9%)	1.33	.72
Miscarriage/patient	4 (12.1%)	4 (9.7%)	1.73	.43

[a] Statistically significant difference: $P < .05$.

Badawy. Clomiphene citrate or letrozole. Fertil Steril 2009.

Bordewijk et al. Eur J Obstet Gynecol Reprod Biol (2020).

Recalculating *p*-values

Table 1: Demographic data of women participated in this study.

		Group A (Tamoxifen) N=100 N (%)	Group B (COCs) N=100 N (%)	P-value	
Age	18-35year	74 (%)	72 (%)	0.750	
	36-45year	26 (%)	28 (%)		
	Mean±SD	33.3±4.7	33.3±4.8		
Level of education	Illiterate	43 (%)	39 (%)	0.135	0.7636
	Some education	45 (%)	46 (%)		
	University	12 (%)	15 (%)		
Husband level of education	Illiterate	20 (%)	14 (%)	0.343	
	Some education	72 (%)	73 (%)		
	University	8 (%)	13 (%)		
No of living children	Mean±SD	3.3±1.6	4.6±1.7	0.502	8.294e-08*
BMI	Mean±SD	31.9±3.2	31.7±3.9	0.243	0.6922

COCs: combined oral contraceptives, BMI: body mass index, SD: standard deviation.

*Changes original significance

Table 3: Effect of treatment on irregular bleeding in both groups.

	Group A (Tamoxifen) N=100		Group B (COCs) N=100		P-value	
Bleeding stopped after treatment	Yes	No	Yes	No	**0.005***	0.0812
	84 (84%)	16 (16%)	92 (92%)	8 (8%)		
No of days required to stop bleeding	**N=84**		**N=92**			
	1-3 day	27 (32%)	1-3 days	13 (14.2%)		
	4-7 day	51 (60.7%)	4-7 days	56 (60.8)	**0.001***	
	8-10 day	6 (7.2%)	8-10 days	16 (17.4%)		
	11-21 day	-	11-21 days	7 (7.6%)		
	Mean±SD	5.03±1.8	Mean±SD	6.5 ±2.5	0.000*	
Percentage of woman did not stop bleeding during treatment	In 3 days	73 (73%)	In 3 days	87 (%)		
	In 7 days	27 (27%)	In 7 days	31 (%)	**0.005***	0.05057 (FET)
	In 10 days	16 (16%)	In 10 days	15 (%)		
	In 21days	-	In 21 days	8 (%)		

*Statistically significant difference, COCs: combined oral contraceptives.

MONASH University

Our results

	Number of Studies	Number of RCTs	Data copying between tables	Last digit (Benford's law)	Baseline data of RCTs	Only even numbers	Statistical mistakes	Retracted	Expression of concern	Period
Dr. Abbas [17]*	263	112	+	NA	+	+	+	8	1	2015 - 2023
Dr. Abd-Elsalaam [16]	163	30	+	NA	+	+	+	7	4	2016 - 2024
Dr. Badawy/AbuHashim [10]	65	35	+	+	+	-	+	24	10	2002 - 2020
Dr. Darwish [18]	28	14	+	+	+	-	+	5	2	1998 - 2023
Dr. Ismail [11]	7	7	+	NA	+	-	+	2	1	2009 - 2019
Dr. Kumar [13]	19	4	NA	NA	+	-	+	4	0	2005 - 2017
Dr. Maged [15]	61	22	+	NA	+	+	+	6	9	2014 - 2023
Dr. Rezk [12]	51	17	+	NA	-	+	+	10	7	2015 - 2020
Dr. Safarinejad** [14]	138	44	NA	NA	NA	+	+	23	7	1996 - 2017
Dr. Shokeir [19]	27	11	+	NA	+	-	NA	4	2	2004 - 2015
Dr. Torky [20]	20	9	+	NA	-	+	+	5	1	2016 - 2021
Total	842	305						98	44	

Reporting findings

- Formal investigation is required to identify scientific misconduct
- Single studies: contact authors, PubPeer, contact journals/publishers
- Groups of studies: peer-reviewed publication, blog posts

- **Formal responses remain slow and inefficient**

Thank you

	Number of Studies	Number of RCTs	Data copying between tables	Last digit (Benford's law)	Baseline data of RCTs	Only even numbers	Statistical mistakes	Retracted	Expression of concern	Period
Dr. Abbas [17]*	263	112	+	NA	+	+	+	8	1	2015 - 2023
Dr. Abd-Elsalaam [16]	163	30	+	NA	+	+	+	7	4	2016 - 2024
Dr. Badawy/AbuHashim [10]	65	35	+	+	+	-	+	24	10	2002 - 2020
Dr. Darwish [18]	28	14	+	+	+	-	+	5	2	1998 - 2023
Dr. Ismail [11]	7	7	+	NA	+	-	+	2	1	2009 - 2019
Dr. Kumar [13]	19	4	NA	NA	+	-	+	4	0	2005 - 2017
Dr. Maged [15]	61	22	+	NA	+	+	+	6	9	2014 - 2023
Dr. Rezk [12]	51	17	+	NA	-	+	+	10	7	2015 - 2020
Dr. Safarinejad** [14]	138	44	NA	NA	NA	+	+	23	7	1996 - 2017
Dr. Shokeir [19]	27	11	+	NA	+	-	NA	4	2	2004 - 2015
Dr. Torky [20]	20	9	+	NA	-	+	+	5	1	2016 - 2021
Total	842	305						98	44	

Jeremy.Nielsen@monash.edu

Mol group

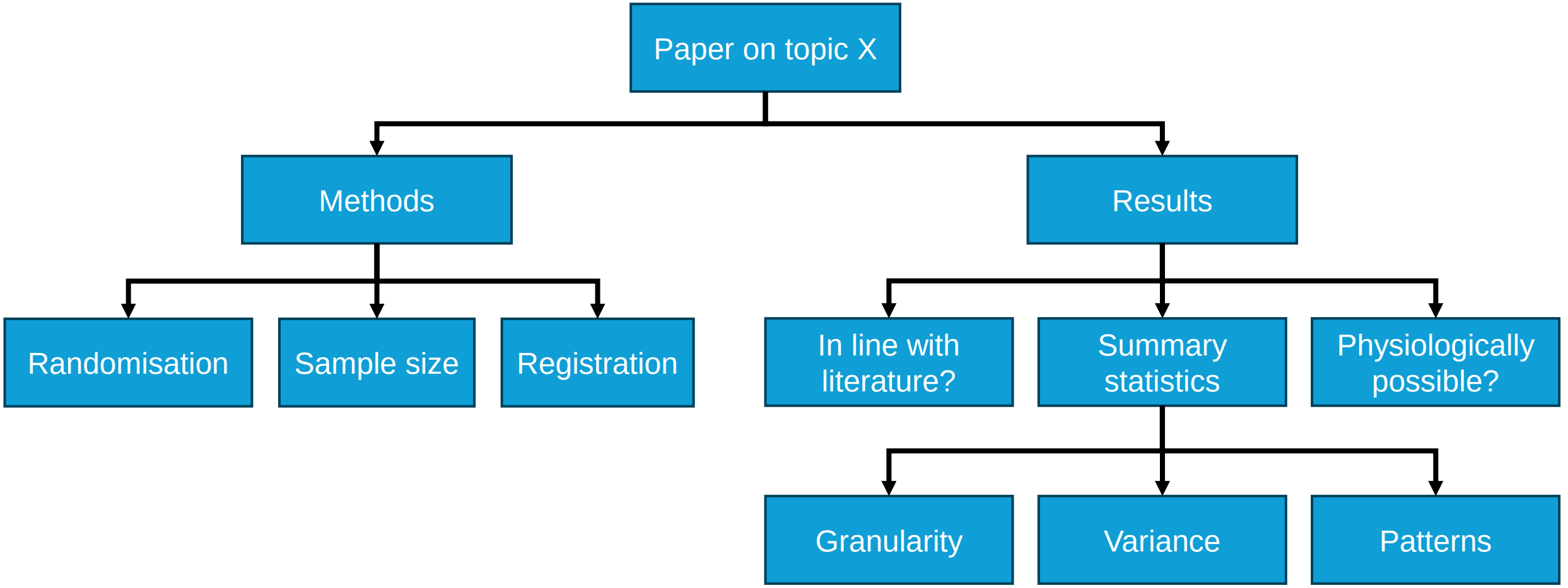
Sue Liu
Siddharth Shivantha
Kelly Zhou
May Linn
Madeline Flanagan

External collaborators

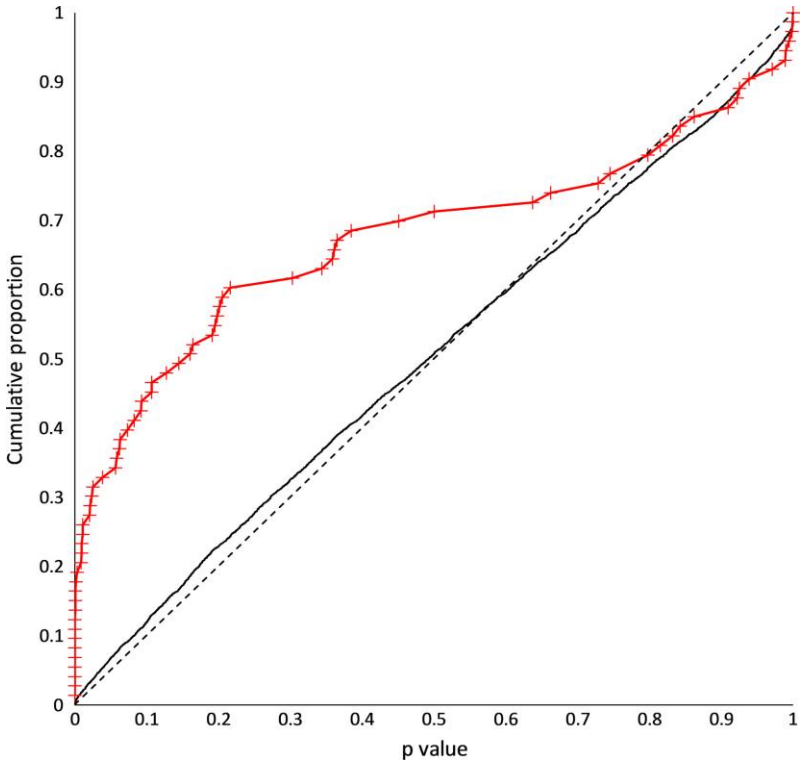
Lyle Gurrin
Jim Thornton
Esmée Bordewijk
Nicholas Brown
Rik van Eekelen
Madelon van Wely

MONASH University
HUDSON INSTITUTE OF MEDICAL RESEARCH
The Ritchie Centre
MERTON COLLEGE OXFORD

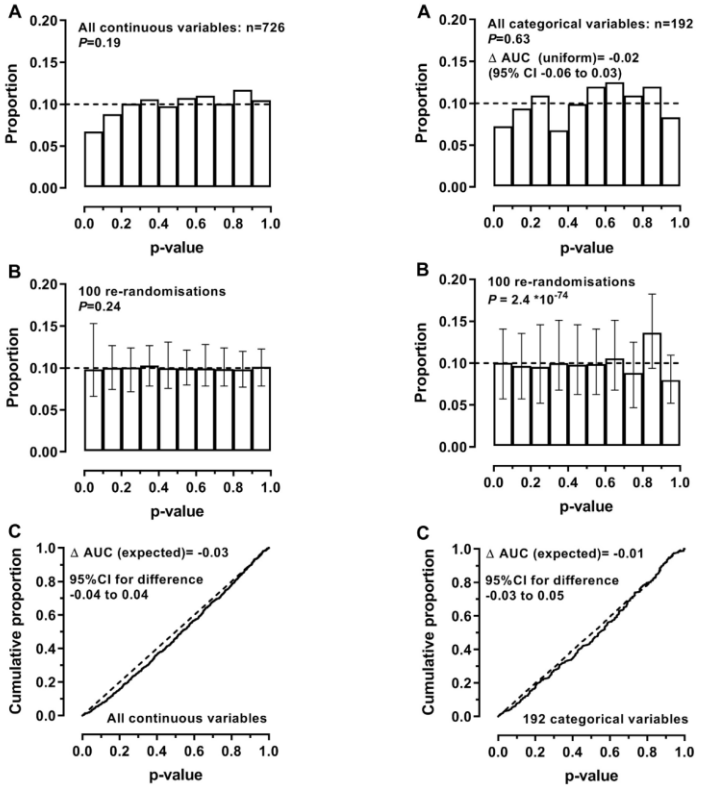
Feasibility assessment



Validation of Carlisle's method



Distributions of p-values from 5015 unretracted trials (black) and 72 retracted trials (red) do not conform to the uniform distribution but are also different to each other. Carlisle. Anaesthesia (2017).



Distribution of p-values is uniform for continuous (Fig. 1, left) but not categorical (Fig. 2, right) variables using data from 13 RCTs by the Auckland group. Categorical variables were uniform, but 100 re-randomisations showed possibility for non-random distributions (Fig. 2B). Bolland et al. J Clin Epidemiol (2019).