Phases of methodological research – and the role of simulations

Georg Heinze,

Anne-Laure Boulesteix, Michael Kammer, Tim Morris, Ian White

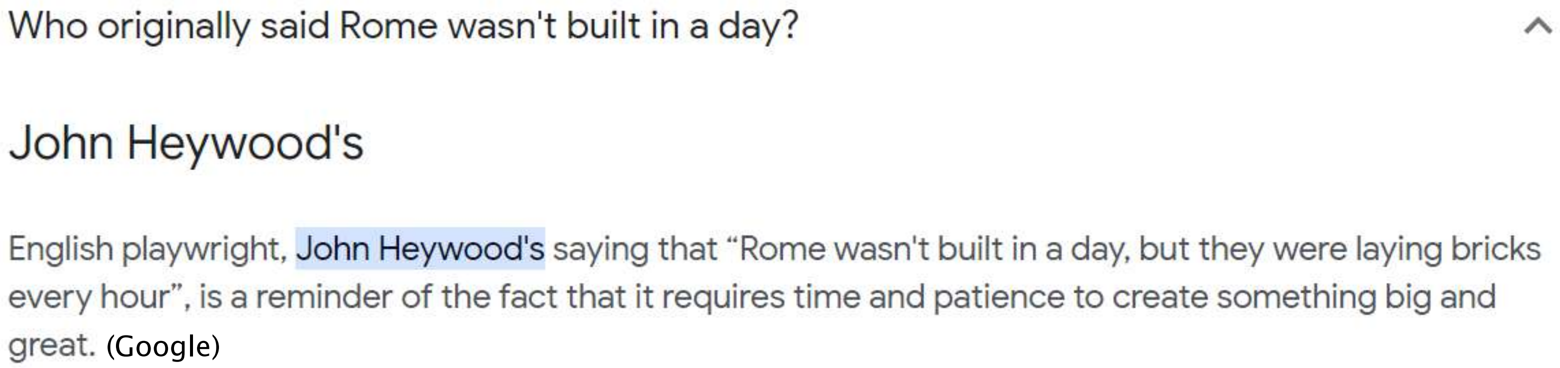
for the Simulation Panel of the STRATOS Initiative

‚Rome wasn‘t built in a day‘

Collosseum 72-80

Pantheon 114-126

St Peter 1506-1626


Who originally said Rome wasn't built in a day?

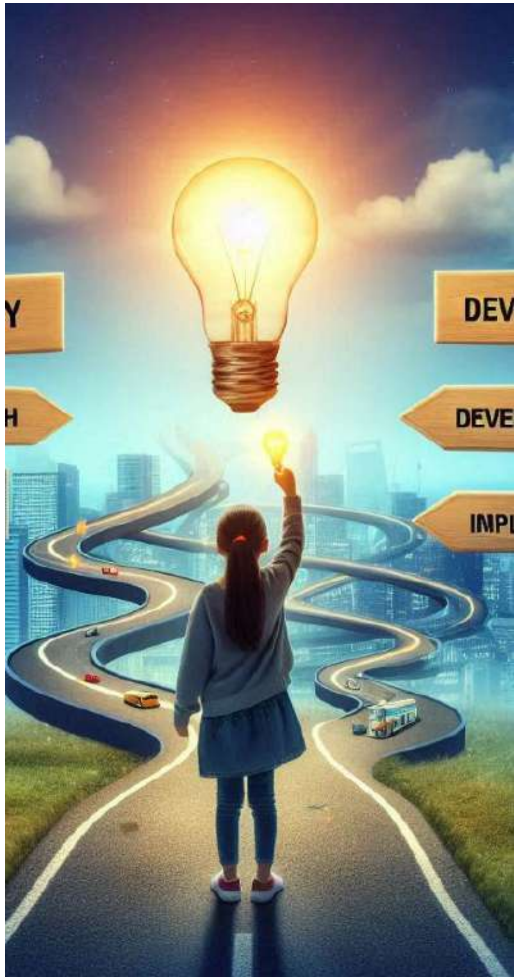
John Heywood's

English playwright, John Heywood's saying that “Rome wasn't built in a day, but they were laying bricks every hour”, is a reminder of the fact that it requires time and patience to create something big and great. (Google)

All pictures Wikipedia or Copilot

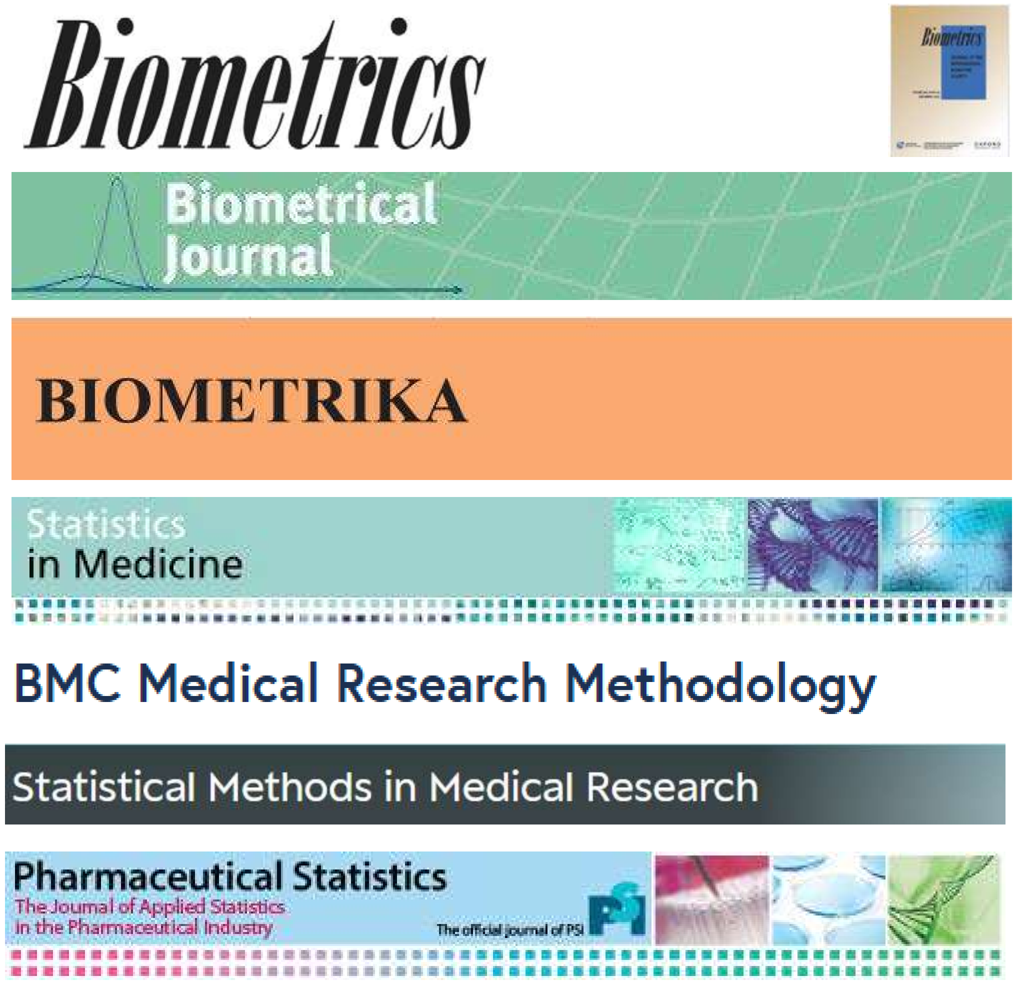
From ideas to trustworthy application: a long way to go

- The goal: „everybody should use it“
- The many obstacles…
- Dead end branches…
- More time than expected…
- Publish or perish..
- The scientific competitors…



All pictures Wikipedia or Copilot

What do we expect from biometrical methods research?


Biometrics
Biometrical Journal
BIOMETRIKA
Statistics in Medicine
BMC Medical Research Methodology
Statistical Methods in Medical Research
Pharmaceutical Statistics
The Journal of Applied Statistics in the Pharmaceutical Industry
The official journal of PSI

- Every issue of these journals is full of newly developed methods

- How many of these methods find their way into routine applications?

All pictures Wikipedia or Copilot

How trustworthy are biometrical methods?

2010

Guidance for industry: adaptive design clinical trials for drugs and biologics
Food and Drug Administration, Washington DC, USA (**2010**)

- Well-understood:
  - Adaptive designs based on:
    - Pretreatment (baseline) data,
    - Blinded interim analysis of aggregate data,
    - Interim results of an outcome unrelated to efficacy,
    - Group sequential designs, unblinded analyses for futility,
    - …

- Less well-understood:
  - Adaptations in dose selection studies,
  - Adaptive randomization based on relative treatment group responses,
  - Adaptation of sample size based on interim effect size estimates,
  - Adaptation of patient population based on treatment-effect estimates
  - …

https://doi.org/10.1016/j.cct.2020.106096

How trustworthy are biometrical methods?

2019

Guidance for industry: adaptive design clinical trials for drugs and biologics
Food and Drug Administration, Washington DC, USA (**2019**)

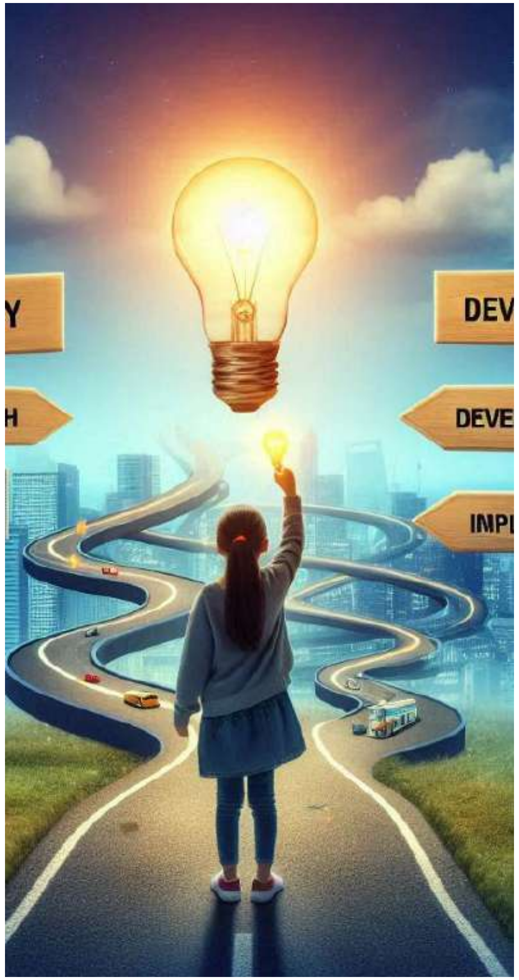
- By 2019, many of the methods were better understood.
- Compared to the initial guidance, this updated guidance provided several examples.
- These motivating examples introduced the advantages of successfully using adaptive designs in the real clinical trials.

https://doi.org/10.1016/j.cct.2020.106096

From ideas to trustworthy application: a typical journey

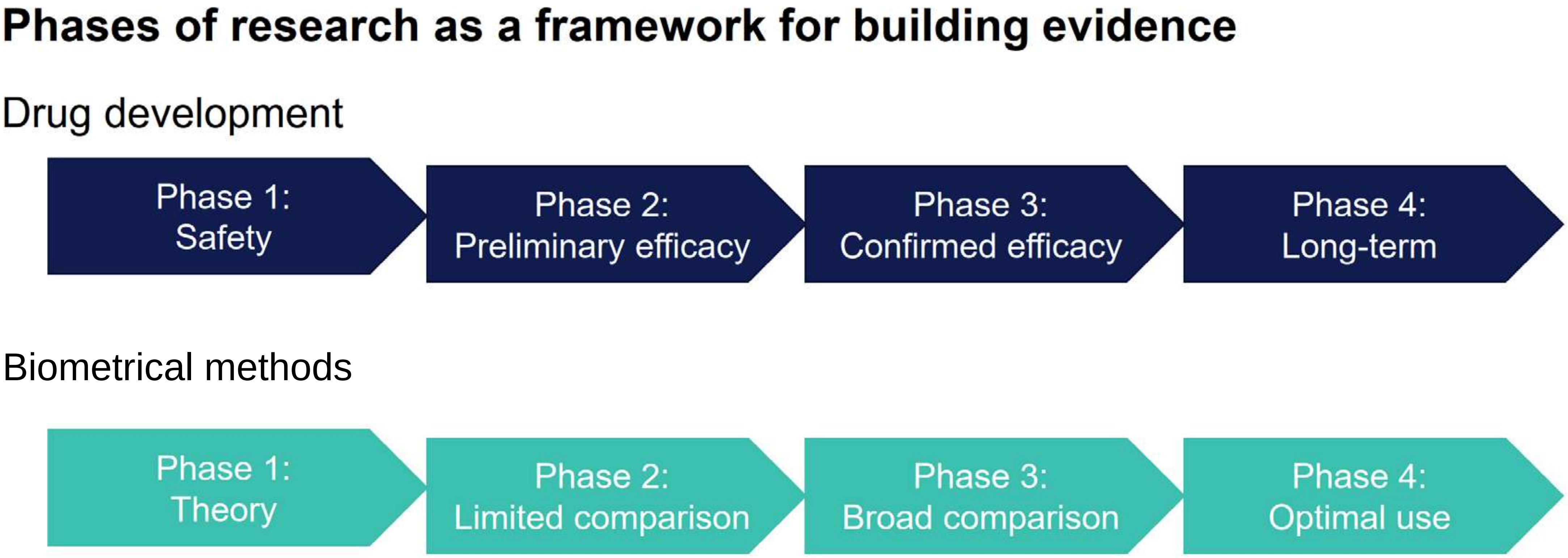
- The methodological problem is identified

- Ideas are generated and described
- A prototype is prepared and compared to alternative methods
- Method is „packaged“, further independent evaluations
- Detailed evaluations, method fully understood, pitfalls and strengths known



All pictures Wikipedia or Copilot

Learning from drug development



Phases of research as a framework for building evidence

Drug development

Phase 1: Safety → Phase 2: Preliminary efficacy → Phase 3: Confirmed efficacy → Phase 4: Long-term

Biometrical methods

Phase 1: Theory → Phase 2: Limited comparison → Phase 3: Broad comparison → Phase 4: Optimal use

Heinze, Boulesteix et al,
Biometrical Journal 2024

The phases of methodological research

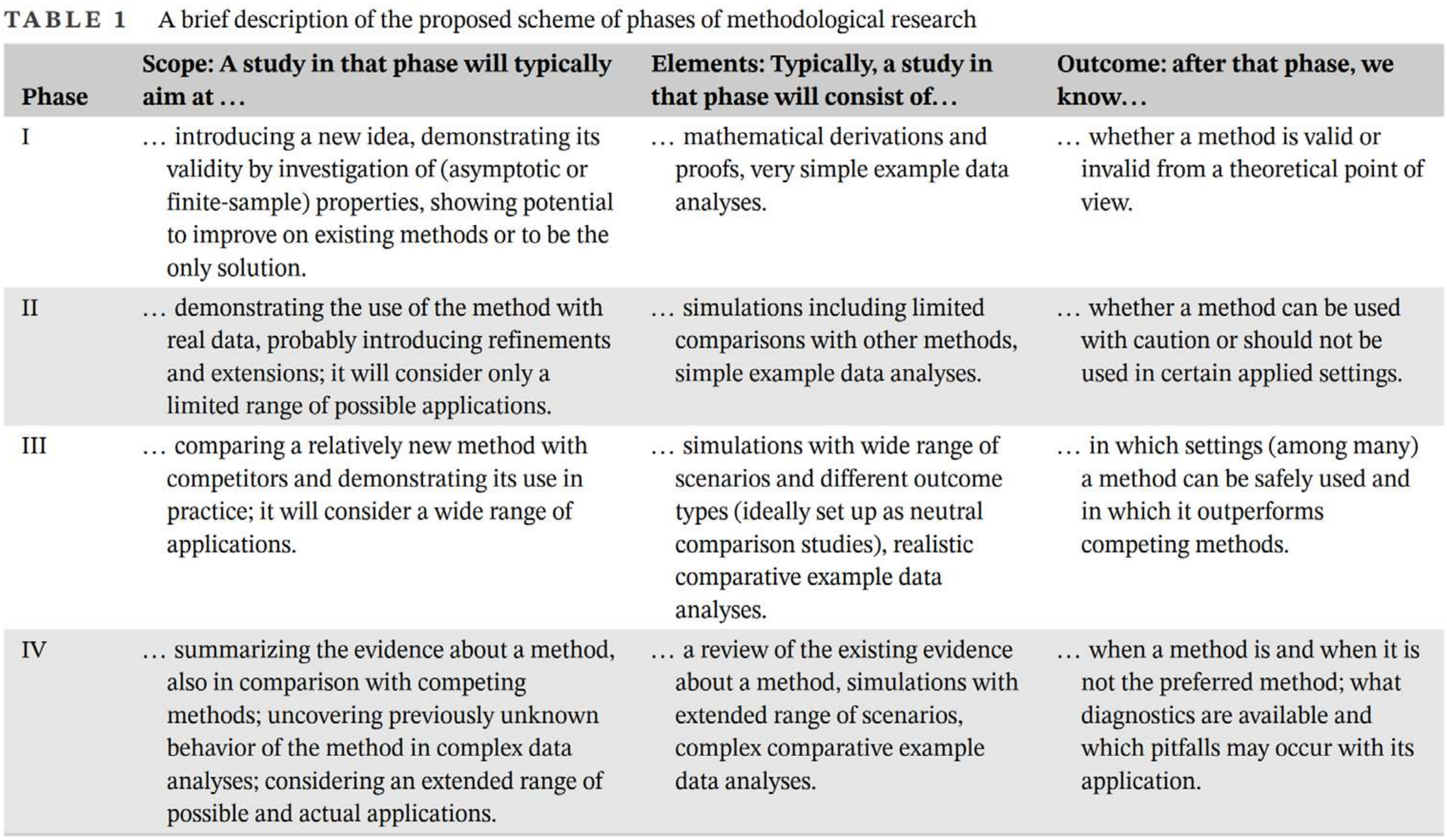


TABLE 1 A brief description of the proposed scheme of phases of methodological research

Phase	Scope: A study in that phase will typically aim at …	Elements: Typically, a study in that phase will consist of…	Outcome: after that phase, we know…
I	… introducing a new idea, demonstrating its validity by investigation of (asymptotic or finite-sample) properties, showing potential to improve on existing methods or to be the only solution.	… mathematical derivations and proofs, very simple example data analyses.	… whether a method is valid or invalid from a theoretical point of view.
II	… demonstrating the use of the method with real data, probably introducing refinements and extensions; it will consider only a limited range of possible applications.	… simulations including limited comparisons with other methods, simple example data analyses.	… whether a method can be used with caution or should not be used in certain applied settings.
III	… comparing a relatively new method with competitors and demonstrating its use in practice; it will consider a wide range of applications.	… simulations with wide range of scenarios and different outcome types (ideally set up as neutral comparison studies), realistic comparative example data analyses.	… in which settings (among many) a method can be safely used and in which it outperforms competing methods.
IV	… summarizing the evidence about a method, also in comparison with competing methods; uncovering previously unknown behavior of the method in complex data analyses; considering an extended range of possible and actual applications.	… a review of the existing evidence about a method, simulations with extended range of scenarios, complex comparative example data analyses.	… when a method is and when it is not the preferred method; what diagnostics are available and which pitfalls may occur with its application.

Heinze, Boulesteix et al,
Biometrical Journal 2024

The role of simulations and synthetic data

- **Phase I**: single ‚toy‘ examples – synthetic data
  - Used to demonstrate how the method works
- **Phase II**: limited range of simulation scenarios
  - To compare the method with (selected) others
  - Still ‚inventor-biased‘

Development

- **Phase III**: broad range of simulation scenarios
  - ‚Neutral‘, wide range of applications in mind

Evaluation, multiple methods

- **Phase IV**: establishing trustworthiness or finding breakdown scenarios
  - May be very wide simulation studies
  - Could be focused on single but likely ‚difficult scenarios‘ (synthetic data sets?)
  - May focus on diagnostics for safe application: when is the method preferred?

Evaluation or review:
one special scenario – many methods
one method – many scenarios

All pictures Wikipedia or Copilot

Example: Firth correction, Phase I

David Firth
University of Warwick

- Firth 1993:
- Enumerated the exact sampling distribution for a toy application example

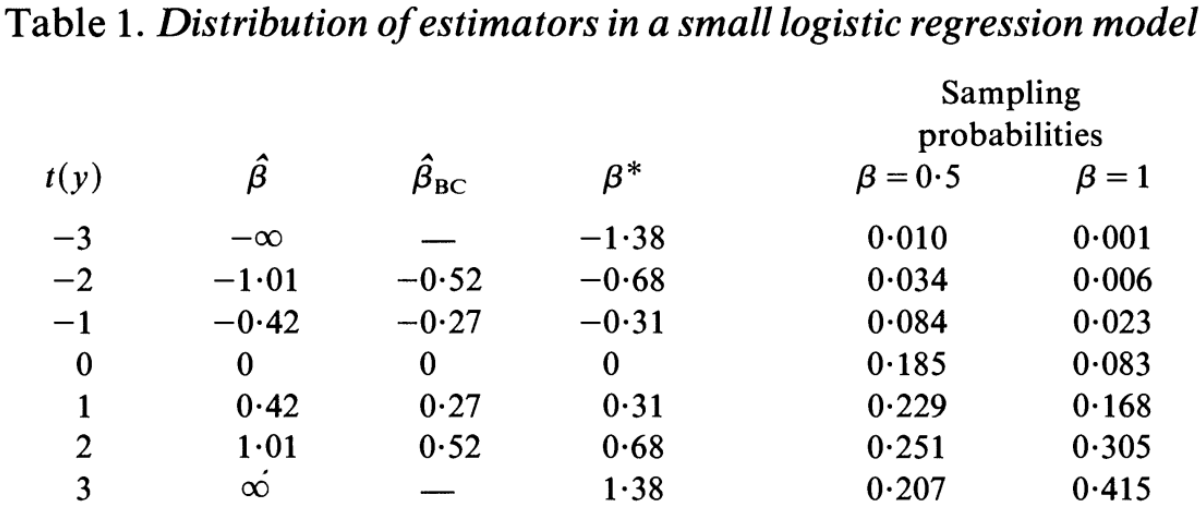


Table 1. *Distribution of estimators in a small logistic regression model*

				Sampling probabilities	
$t(y)$	$\hat\beta$	$\hat\beta_{BC}$	β^*	$\beta = 0{\cdot}5$	$\beta = 1$
−3	$-\infty$	—	−1·38	0·010	0·001
−2	−1·01	−0·52	−0·68	0·034	0·006
−1	−0·42	−0·27	−0·31	0·084	0·023
0	0	0	0	0·185	0·083
1	0·42	0·27	0·31	0·229	0·168
2	1·01	0·52	0·68	0·251	0·305
3	∞	—	1·38	0·207	0·415

- No other empirical data presented

All pictures Wikipedia or Copilot

Example: Firth correction, Phase II

- Heinze and Schemper 2002:

- Simple simulations:
  assuming independence of risk factors,
  only dichotomous risk factors,
  strong effects only,
  limited scope of sample sizes,
  ‚edgy‘ scenarios

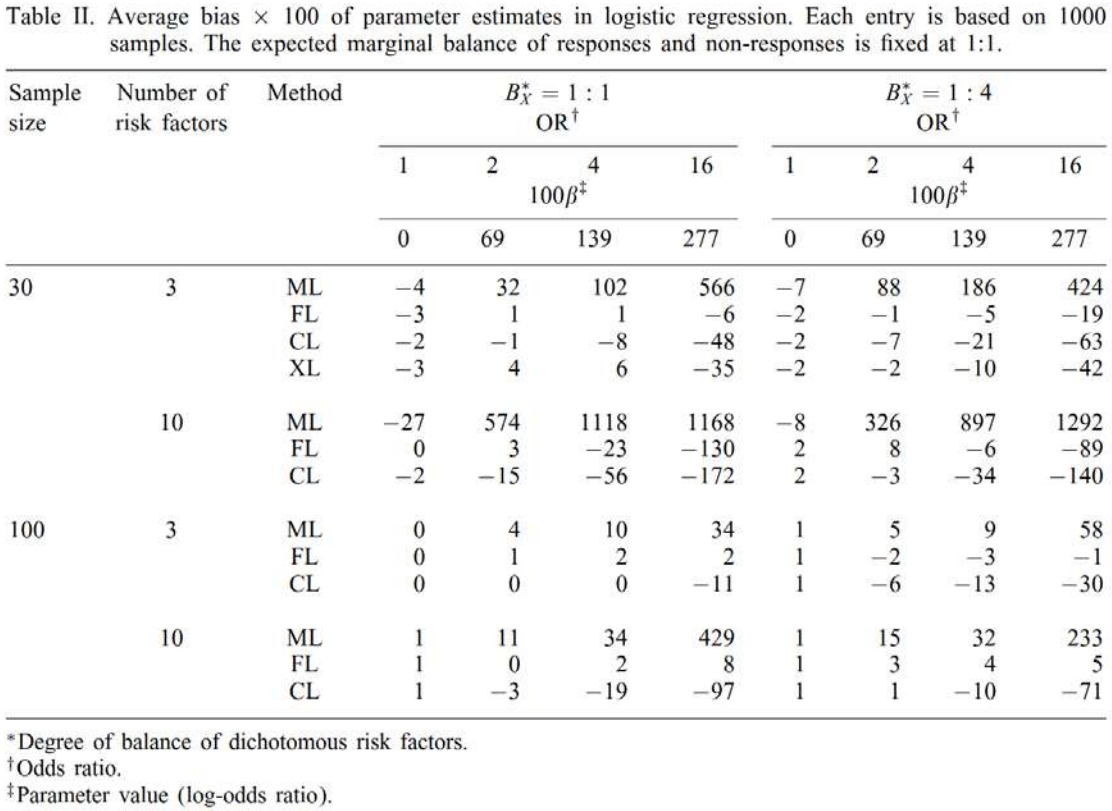


Table II. Average bias × 100 of parameter estimates in logistic regression. Each entry is based on 1000 samples. The expected marginal balance of responses and non-responses is fixed at 1:1.

Sample size	Number of risk factors	Method	$B_X^* = 1:1$ OR[†]				$B_X^* = 1:4$ OR[†]			
			1	2	4	16	1	2	4	16
			100β[‡]				100β[‡]			
			0	69	139	277	0	69	139	277
30	3	ML	−4	32	102	566	−7	88	186	424
		FL	−3	1	1	−6	−2	−1	−5	−19
		CL	−2	−1	−8	−48	−2	−7	−21	−63
		XL	−3	4	6	−35	−2	−2	−10	−42
	10	ML	−27	574	1118	1168	−8	326	897	1292
		FL	0	3	−23	−130	2	8	−6	−89
		CL	−2	−15	−56	−172	2	−3	−34	−140
100	3	ML	0	4	10	34	1	5	9	58
		FL	0	1	2	2	1	−2	−3	−1
		CL	0	0	0	−11	1	−6	−13	−30
	10	ML	1	11	34	429	1	15	32	233
		FL	1	0	2	8	1	3	4	5
		CL	1	−3	−19	−97	1	1	−10	−71

[*] Degree of balance of dichotomous risk factors.
[†] Odds ratio.
[‡] Parameter value (log-odds ratio).

All pictures Wikipedia or Copilot

Example: Firth correction, Phase III

- Van Smeden et al, 2016
- Highly factorial design of simulation study
  - Focus on Events per Variable
  - Realistic effect sizes
  - Small number of covariates
  - Simple correlation patterns
- High-level summary of results

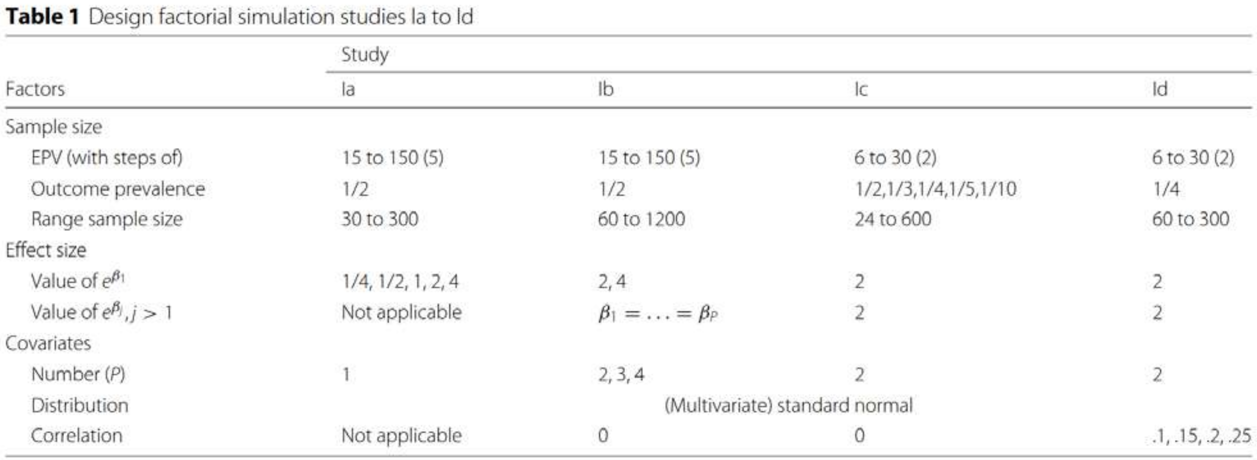


Table 1 Design factorial simulation studies Ia to Id

	Study			
Factors	Ia	Ib	Ic	Id
Sample size				
EPV (with steps of)	15 to 150 (5)	15 to 150 (5)	6 to 30 (2)	6 to 30 (2)
Outcome prevalence	1/2	1/2	1/2,1/3,1/4,1/5,1/10	1/4
Range sample size	30 to 300	60 to 1200	24 to 600	60 to 300
Effect size				
Value of e^{β_1}	1/4, 1/2, 1, 2, 4	2, 4	2	2
Value of $e^{\beta_j}, j > 1$	Not applicable	$\beta_1 = \ldots = \beta_P$	2	2
Covariates				
Number (*P*)	1	2, 3, 4	2	2
Distribution		(Multivariate) standard normal		
Correlation	Not applicable	0	0	.1, .15, .2, .25

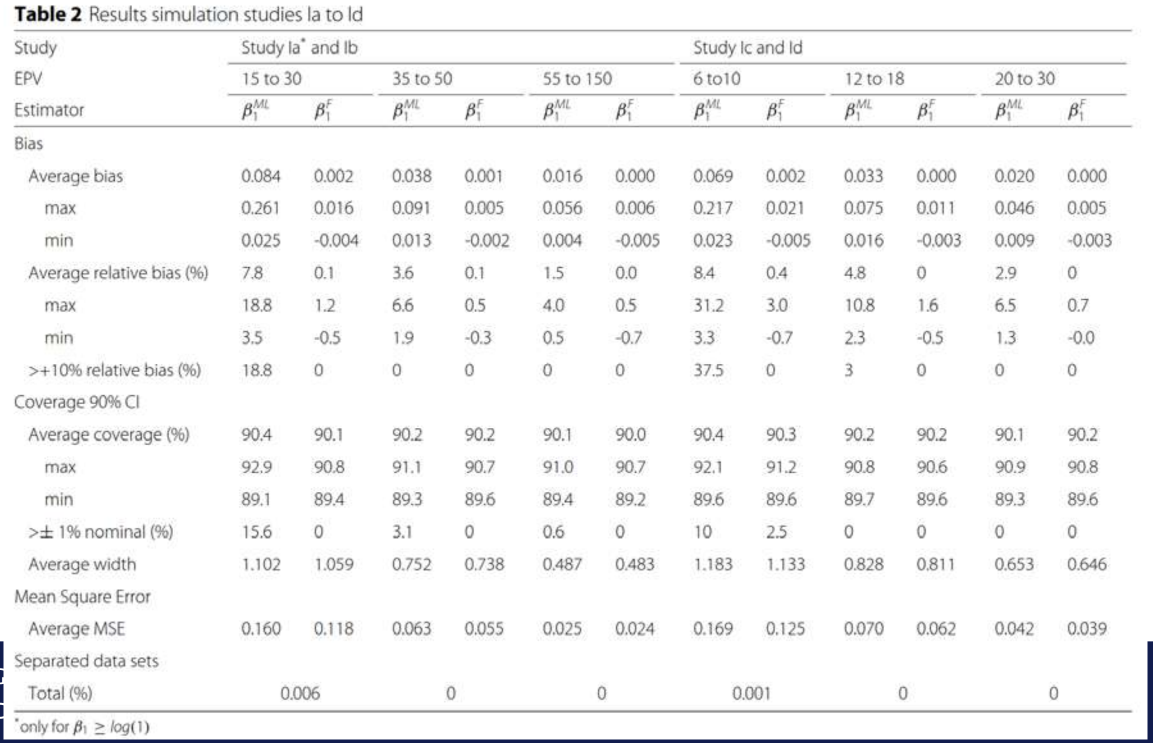


Table 2 Results simulation studies Ia to Id

Study	Study Ia* and Ib						Study Ic and Id					
EPV	15 to 30		35 to 50		55 to 150		6 to10		12 to 18		20 to 30	
Estimator	β_1^{ML}	β_1^{F}	β_1^{ML}	β_1^{F}	β_1^{ML}	β_1^{F}	β_1^{ML}	β_1^{F}	β_1^{ML}	β_1^{F}	β_1^{ML}	β_1^{F}
Bias												
Average bias	0.084	0.002	0.038	0.001	0.016	0.000	0.069	0.002	0.033	0.000	0.020	0.000
max	0.261	0.016	0.091	0.005	0.056	0.006	0.217	0.021	0.075	0.011	0.046	0.005
min	0.025	-0.004	0.013	-0.002	0.004	-0.005	0.023	-0.005	0.016	-0.003	0.009	-0.003
Average relative bias (%)	7.8	0.1	3.6	0.1	1.5	0.0	8.4	0.4	4.8	0	2.9	0
max	18.8	1.2	6.6	0.5	4.0	0.5	31.2	3.0	10.8	1.6	6.5	0.7
min	3.5	-0.5	1.9	-0.3	0.5	-0.7	3.3	-0.7	2.3	-0.5	1.3	-0.0
>+10% relative bias (%)	18.8	0	0	0	0	0	37.5	0	3	0	0	0
Coverage 90% CI												
Average coverage (%)	90.4	90.1	90.2	90.2	90.1	90.0	90.4	90.3	90.2	90.2	90.1	90.2
max	92.9	90.8	91.1	90.7	91.0	90.7	92.1	91.2	90.8	90.6	90.9	90.8
min	89.1	89.4	89.3	89.6	89.4	89.2	89.6	89.6	89.7	89.6	89.3	89.6
>± 1% nominal (%)	15.6	0	3.1	0	0.6	0	10	2.5	0	0	0	0
Average width	1.102	1.059	0.752	0.738	0.487	0.483	1.183	1.133	0.828	0.811	0.653	0.646
Mean Square Error												
Average MSE	0.160	0.118	0.063	0.055	0.025	0.024	0.169	0.125	0.070	0.062	0.042	0.039
Separated data sets												
Total (%)	0.006		0		0		0.001		0		0	

*only for $\beta_1 \geq log(1)$

All pictures Wikipedia or Copilot

Example: Firth correction: Phase III, and back to Phases I-II

- Puhr et al, 2017

- Compared various methods to deal with separation in prediction setting
  - 9 main scenarios
  - Mixed types of covariates (realistic)
  - Realistic effect sizes
  - Realistic sample sizes

- Introduced two new methods to alleviate known problems with Firth correction: FLIC and FLAC

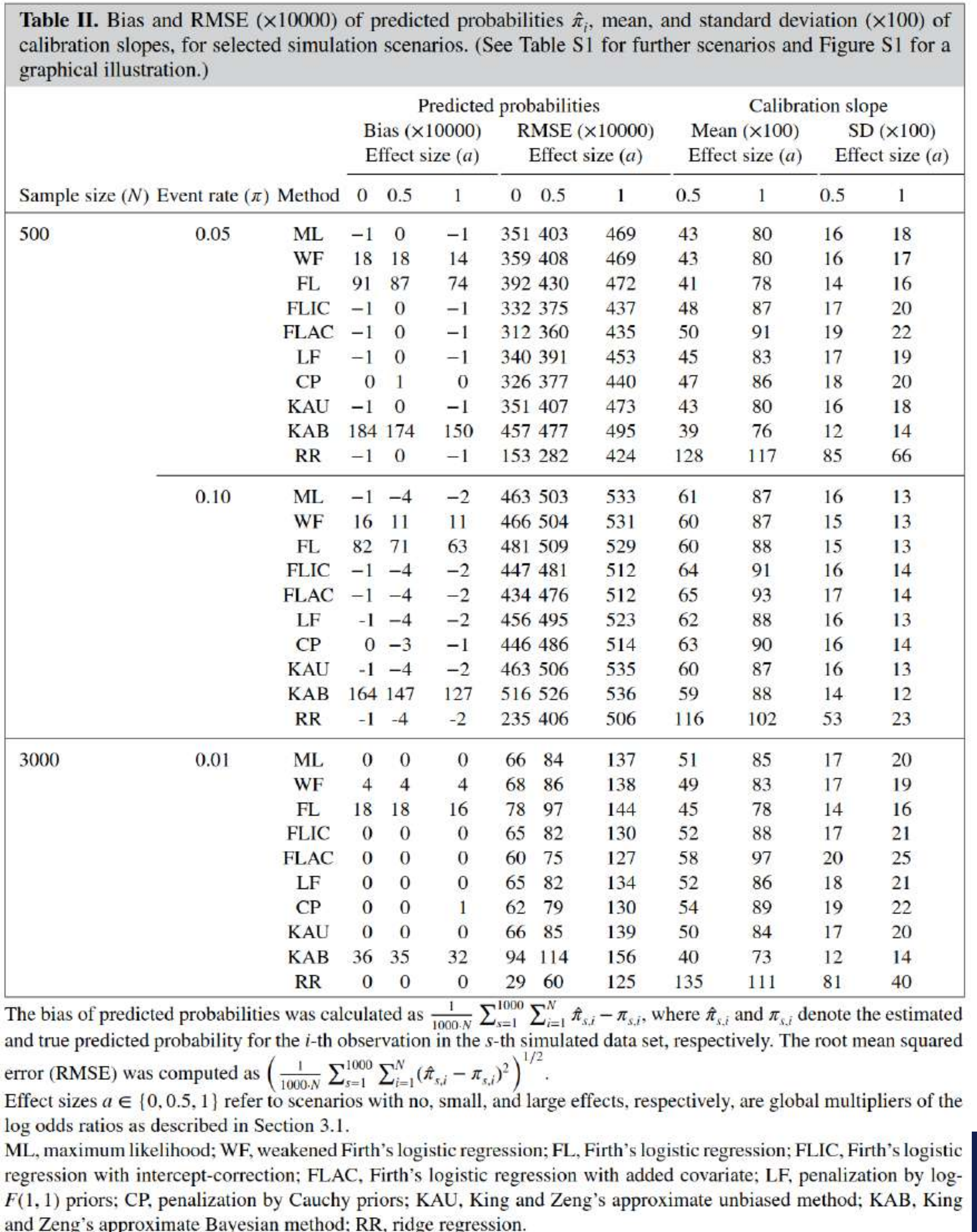


Table II. Bias and RMSE (×10000) of predicted probabilities $\hat{\pi}_i$, mean, and standard deviation (×100) of calibration slopes, for selected simulation scenarios. (See Table S1 for further scenarios and Figure S1 for a graphical illustration.)

			Predicted probabilities						Calibration slope			
			Bias (×10000)			RMSE (×10000)			Mean (×100)		SD (×100)	
			Effect size (a)			Effect size (a)			Effect size (a)		Effect size (a)	
Sample size (N)	Event rate (π)	Method	0	0.5	1	0	0.5	1	0.5	1	0.5	1
500	0.05	ML	−1	0	−1	351	403	469	43	80	16	18
		WF	18	18	14	359	408	469	43	80	16	17
		FL	91	87	74	392	430	472	41	78	14	16
		FLIC	−1	0	−1	332	375	437	48	87	17	20
		FLAC	−1	0	−1	312	360	435	50	91	19	22
		LF	−1	0	−1	340	391	453	45	83	17	19
		CP	0	1	0	326	377	440	47	86	18	20
		KAU	−1	0	−1	351	407	473	43	80	16	18
		KAB	184	174	150	457	477	495	39	76	12	14
		RR	−1	0	−1	153	282	424	128	117	85	66
	0.10	ML	−1	−4	−2	463	503	533	61	87	16	13
		WF	16	11	11	466	504	531	60	87	15	13
		FL	82	71	63	481	509	529	60	88	15	13
		FLIC	−1	−4	−2	447	481	512	64	91	16	14
		FLAC	−1	−4	−2	434	476	512	65	93	17	14
		LF	-1	−4	−2	456	495	523	62	88	16	13
		CP	0	−3	−1	446	486	514	63	90	16	14
		KAU	-1	−4	−2	463	506	535	60	87	16	13
		KAB	164	147	127	516	526	536	59	88	14	12
		RR	-1	-4	-2	235	406	506	116	102	53	23
3000	0.01	ML	0	0	0	66	84	137	51	85	17	20
		WF	4	4	4	68	86	138	49	83	17	19
		FL	18	18	16	78	97	144	45	78	14	16
		FLIC	0	0	0	65	82	130	52	88	17	21
		FLAC	0	0	0	60	75	127	58	97	20	25
		LF	0	0	0	65	82	134	52	86	18	21
		CP	0	0	1	62	79	130	54	89	19	22
		KAU	0	0	0	66	85	139	50	84	17	20
		KAB	36	35	32	94	114	156	40	73	12	14
		RR	0	0	0	29	60	125	135	111	81	40

The bias of predicted probabilities was calculated as $\frac{1}{1000 \cdot N} \sum_{s=1}^{1000} \sum_{i=1}^{N} \hat{\pi}_{s,i} - \pi_{s,i}$, where $\hat{\pi}_{s,i}$ and $\pi_{s,i}$ denote the estimated and true predicted probability for the i-th observation in the s-th simulated data set, respectively. The root mean squared error (RMSE) was computed as $\left(\frac{1}{1000 \cdot N} \sum_{s=1}^{1000} \sum_{i=1}^{N} (\hat{\pi}_{s,i} - \pi_{s,i})^2\right)^{1/2}$.

Effect sizes $a \in \{0, 0.5, 1\}$ refer to scenarios with no, small, and large effects, respectively, are global multipliers of the log odds ratios as described in Section 3.1.

ML, maximum likelihood; WF, weakened Firth's logistic regression; FL, Firth's logistic regression; FLIC, Firth's logistic regression with intercept-correction; FLAC, Firth's logistic regression with added covariate; LF, penalization by log-$F(1, 1)$ priors; CP, penalization by Cauchy priors; KAU, King and Zeng's approximate unbiased method; KAB, King and Zeng's approximate Bayesian method; RR, ridge regression.

Example: Firth correction: Phase IV

- Mansournia et al, 2018:
- Team of authors from ‚different camps‘
- Review of the problem, differences in software results, review of solutions, balanced discussion of solutions

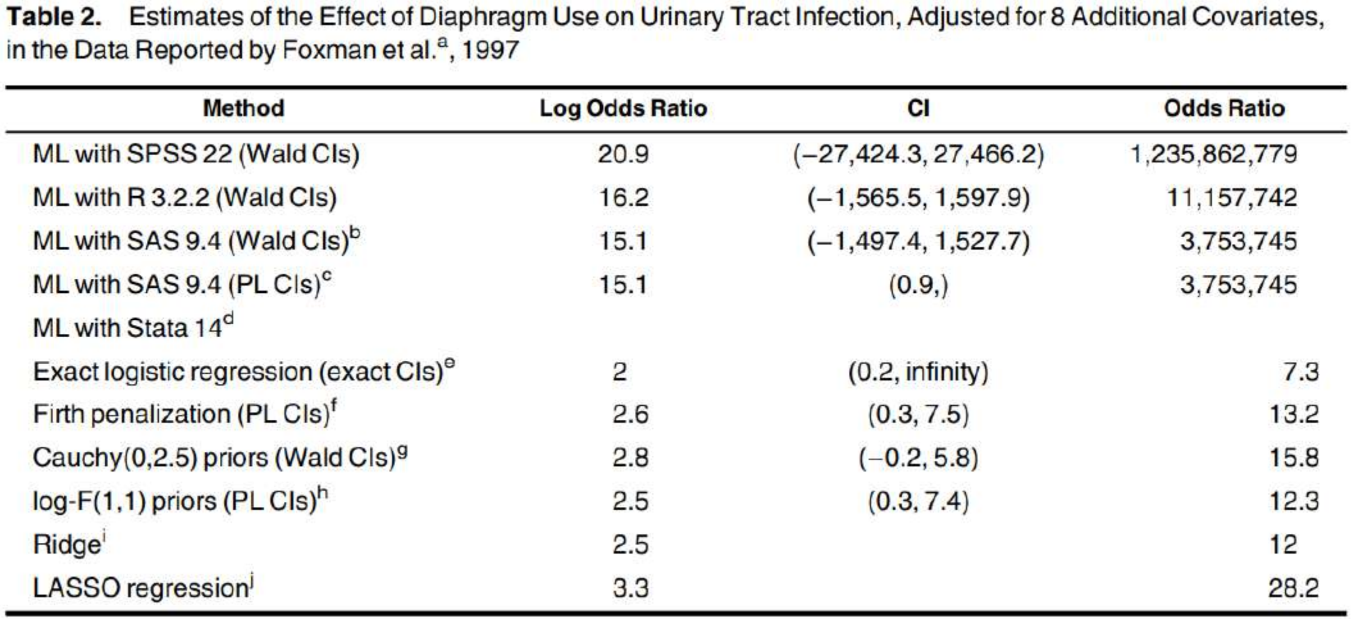


Table 2. Estimates of the Effect of Diaphragm Use on Urinary Tract Infection, Adjusted for 8 Additional Covariates, in the Data Reported by Foxman et al.[a], 1997

Method	Log Odds Ratio	CI	Odds Ratio
ML with SPSS 22 (Wald CIs)	20.9	(−27,424.3, 27,466.2)	1,235,862,779
ML with R 3.2.2 (Wald CIs)	16.2	(−1,565.5, 1,597.9)	11,157,742
ML with SAS 9.4 (Wald CIs)[b]	15.1	(−1,497.4, 1,527.7)	3,753,745
ML with SAS 9.4 (PL CIs)[c]	15.1	(0.9,)	3,753,745
ML with Stata 14[d]			
Exact logistic regression (exact CIs)[e]	2	(0.2, infinity)	7.3
Firth penalization (PL CIs)[f]	2.6	(0.3, 7.5)	13.2
Cauchy(0,2.5) priors (Wald CIs)[g]	2.8	(−0.2, 5.8)	15.8
log-F(1,1) priors (PL CIs)[h]	2.5	(0.3, 7.4)	12.3
Ridge[i]	2.5		12
LASSO regression[j]	3.3		28.2

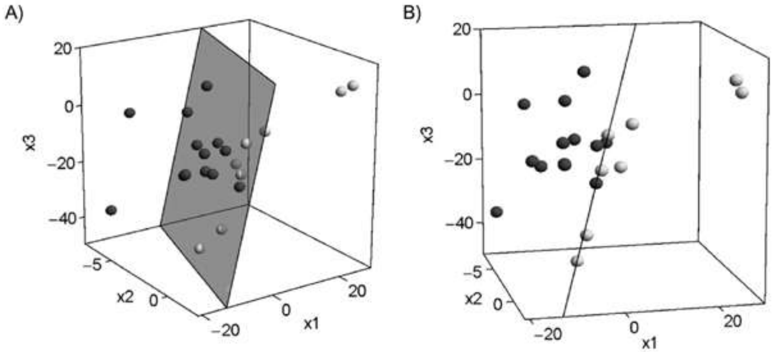
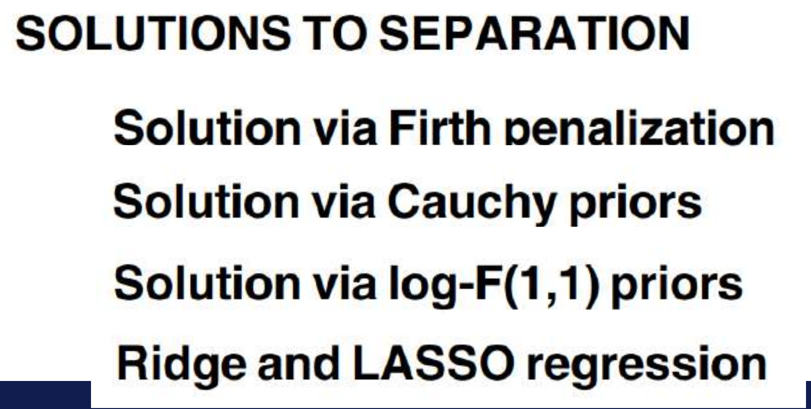


Figure 1. Illustration of data separation for the data from Potter (11), 2005. The axes correspond to the 3 covariates. Treatment success is marked in black and failure in gray. Plots (A) and (B) differ only in the angle of view. The data are an example of quasicomplete separation (i.e., there is a plane (with equation $-112.3x_1 - 165.3x_2 + 21.02x_3 = 5.4$) that separates data points with different outcomes but with observations of both outcomes lying exactly on the plane).



SOLUTIONS TO SEPARATION

Solution via Firth penalization

Solution via Cauchy priors

Solution via log-F(1,1) priors

Ridge and LASSO regression

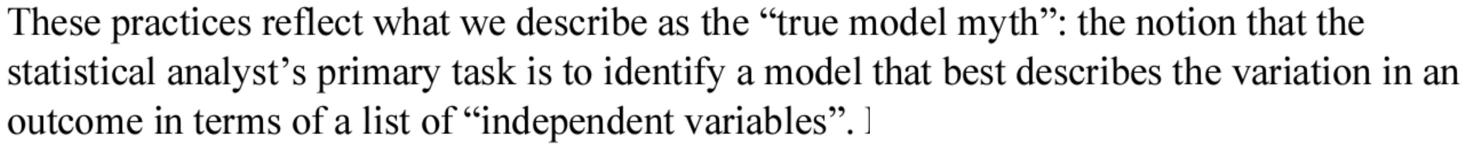
The role of simulations

- Can play a critical role in each phase
  - Phase I: very simple, illustrate feasibility
  - Phase II: limited range of scenarios, controlled conditions
  - Phase III: broad comparisons, neutrality
  - Phase IV: no must, but may reveal weaknesses or demonstrate robustness
- Simulations rarely used alone, but they complement:
  - Theoretical analyses by empiricial evidence
  - Real-world data analyses revealing aspects of application

All pictures Wikipedia or Copilot

Pitfalls in simulation studies (all phases)

- Too strong belief in the ‚true model‘:



These practices reflect what we describe as the “true model myth”: the notion that the statistical analyst’s primary task is to identify a model that best describes the variation in an outcome in terms of a list of “independent variables”.

Carlin and Moreno-Betancur, arXiv 2024
upcoming in Stat Med (2025)

- What Carlin and Moreno-Betancur describe also applies to simulation studies:
  - Don‘t believe that data is ever generated by a ‚model‘ with independent Gaussian errors
  - Phase II studies often exhibit a clear ‚winner‘ method:
    the method that magically captures features of the data generating mechanism
  - Should we move on? Towards methods comparison studies!
    Separate data generation from data analysis in simulation studies

Don‘t expect anyone to be able to build Rome in one day!

- Authors of methods research should clearly disclose the phase they‘re contributing to!

- For Phase I: do not ask authors to prove that their new method works in all hypothetical scenario; allow ‚high-risk‘ methods
- For Phase II: reduce the risk: comparison included? Data example?
- For Phase III-IV: specifically check neutrality and broadness of comparisons. Realistic data example?
- For Phase IV: is it clear when the method is to preferred over others and when not?

It is a long way from ideas to trustworthy application!

- Journals should
  - encourage authors to clearly disclose the phase they‘re contributing to!

- Funding agencies should
  - not accept proposals that claim to cover all phases
    (from invention to implementation)!
  - accept good proposals that aim to evaluate existing methods!

- PhD evaluators, tenure track evaluators should
  - see work in context to the phase
    - all phases are important, no work needs to cover more than one phase
  - consider neutral comparison studies as valuable scientific contributions
    - appreciate that they are difficult to design and conduct (not just ‚bigger simulation studies‘)

- *Accept that research needs time for development and evaluation!*

Also Vienna wasn‘t built in a day

1359-1439
(tower)

1696-1705, 1742-1749

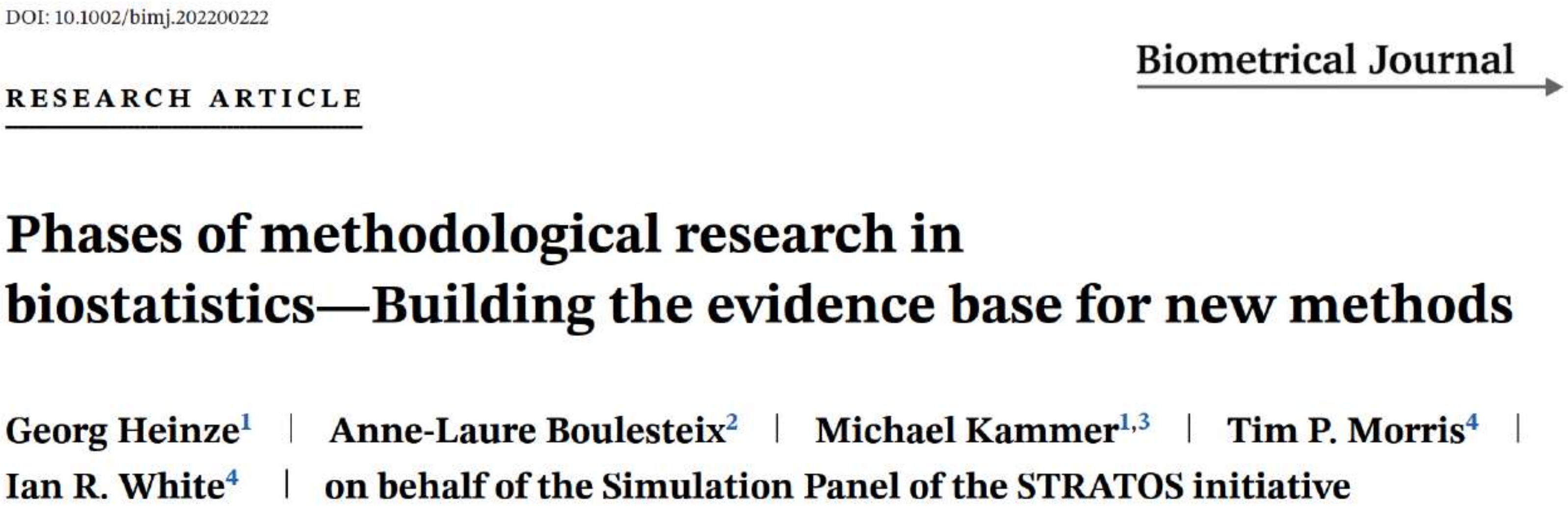
1896-1897

1964-1994

2020-2026*

All pictures Wikipedia or Copilot

Reference



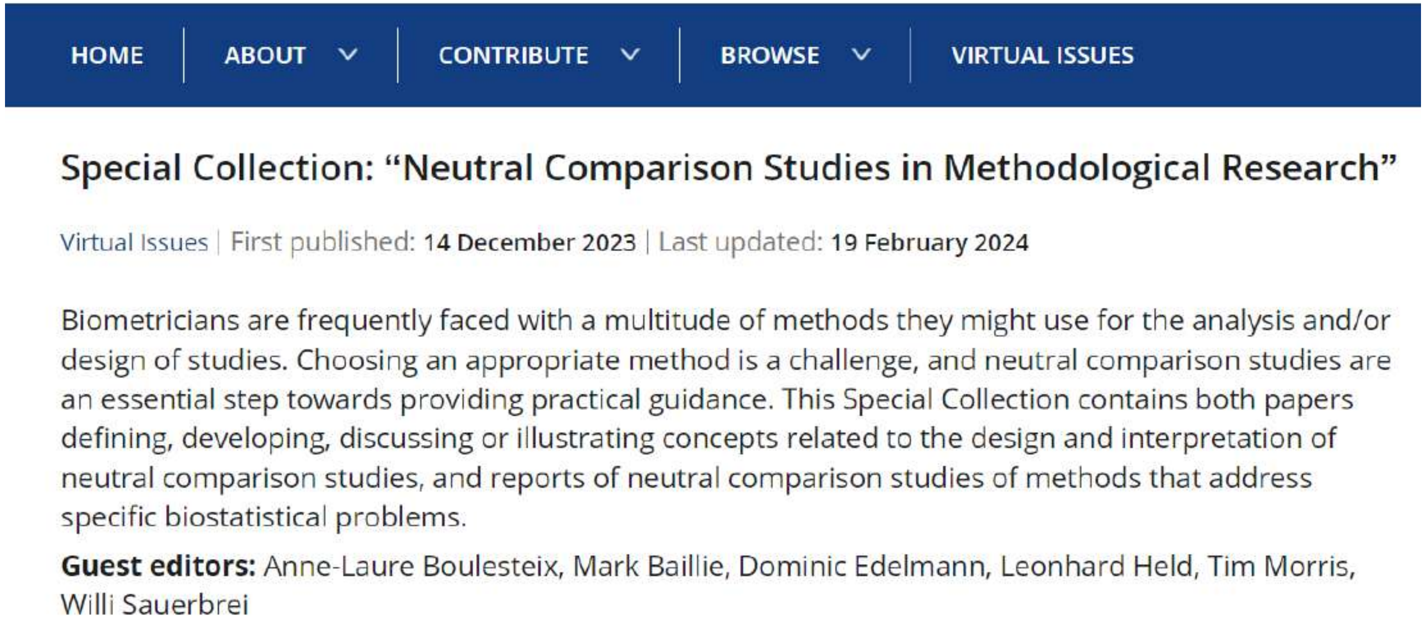
DOI: 10.1002/bimj.202200222

RESEARCH ARTICLE

Biometrical Journal

Phases of methodological research in biostatistics—Building the evidence base for new methods

Georg Heinze[1] | Anne-Laure Boulesteix[2] | Michael Kammer[1,3] | Tim P. Morris[4] | Ian R. White[4] | on behalf of the Simulation Panel of the STRATOS initiative



Special Collection: "Neutral Comparison Studies in Methodological Research"

Virtual Issues | First published: 14 December 2023 | Last updated: 19 February 2024

Biometricians are frequently faced with a multitude of methods they might use for the analysis and/or design of studies. Choosing an appropriate method is a challenge, and neutral comparison studies are an essential step towards providing practical guidance. This Special Collection contains both papers defining, developing, discussing or illustrating concepts related to the design and interpretation of neutral comparison studies, and reports of neutral comparison studies of methods that address specific biostatistical problems.

Guest editors: Anne-Laure Boulesteix, Mark Baillie, Dominic Edelmann, Leonhard Held, Tim Morris, Willi Sauerbrei

All pictures Wikipedia or Copilot