Replicability Issues in Epidemiologic Research

Annibale Biggeri

Department of Cardiac, Thoracic, Vascular Sciences and Public Health

Unit of Biostatistics, Epidemiology and Public Health

University of Padua

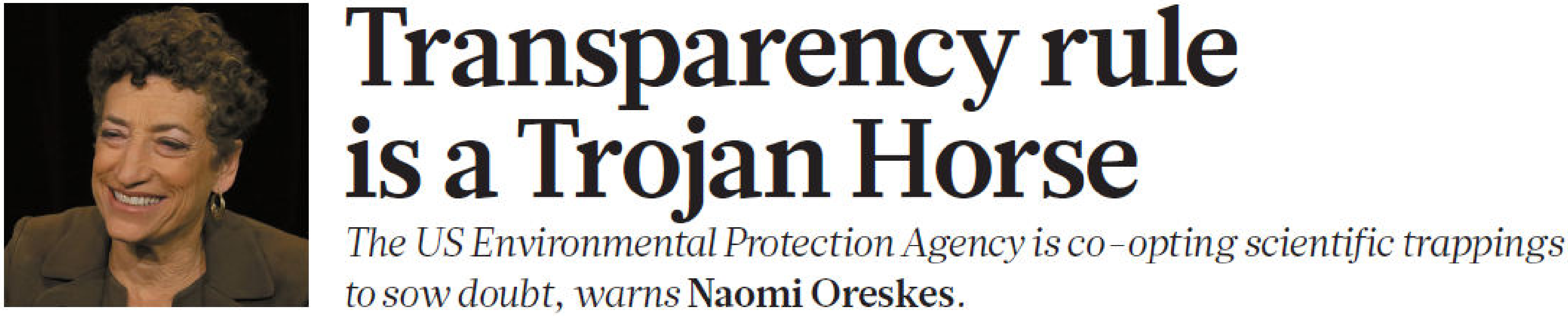
annibale.biggeri@unipd.it

Motivating example

On April 24, 2018, U.S. Environmental Protection Agency (EPA) **Administrator** Scott Pruitt signed a proposed rule to strengthen the science used in EPA regulations. This is justified recognizing a “replication crisis”— in EPA wording “a significant proportion of published research may not be reproducible”. This position coupled with the appointment of Louis Cox as chairman EPA’s Clean Air Scientific Advisory Committee (CASAC) aimed at returning CASAC “to its original scientific mission and may help restore scientific integrity and political accountability that are essential to effective **environmental policy**”. This because in the past “its members have strayed from their mandate to advise on scientific questions and become vociferous advocates for (or opponents of) certain policies”.

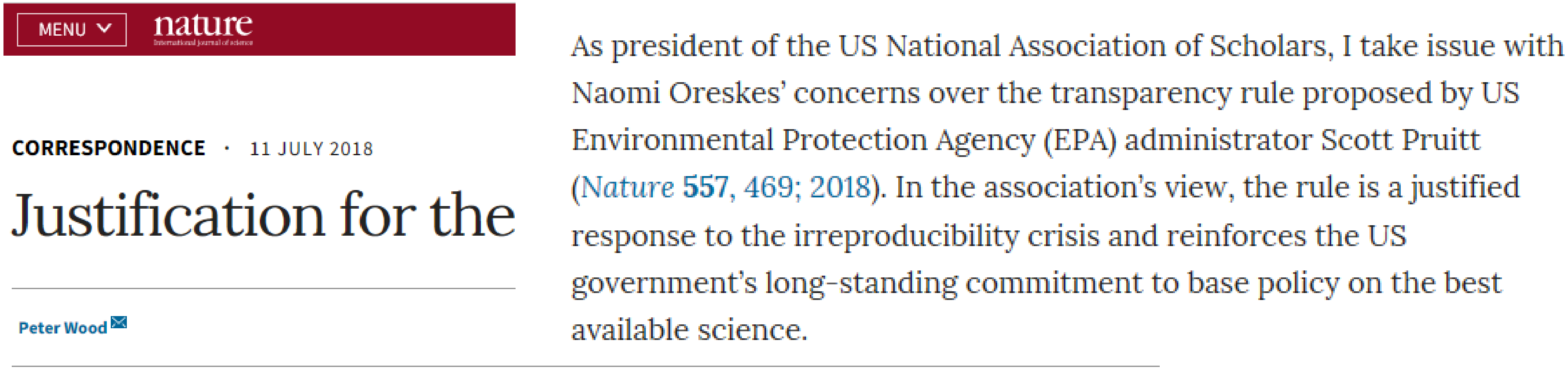
Playing Replicability

It is clear that the replicability issue is used as a strategy to weaken the results of environmental epidemiology on health effects of air pollution (Oreskes Nature 2018).



Transparency rule is a Trojan Horse

The US Environmental Protection Agency is co-opting scientific trappings to sow doubt, warns **Naomi Oreskes**.



CORRESPONDENCE · 11 JULY 2018

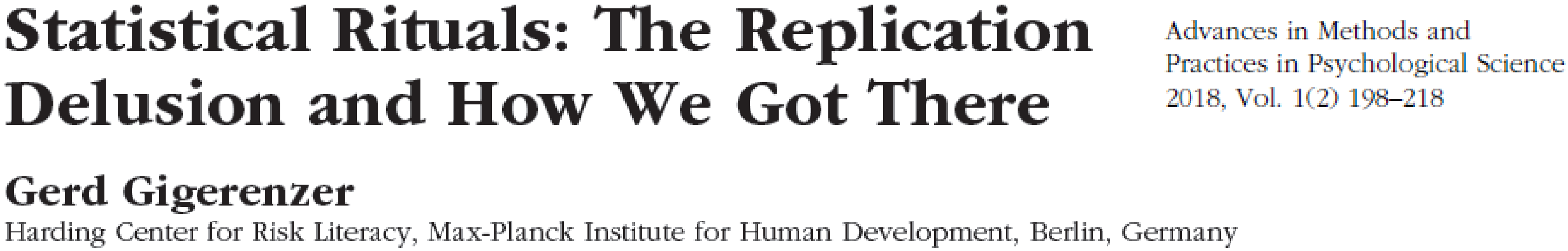
Justification for the

Peter Wood ✉

As president of the US National Association of Scholars, I take issue with Naomi Oreskes' concerns over the transparency rule proposed by US Environmental Protection Agency (EPA) administrator Scott Pruitt (*Nature* 557, 469; 2018). In the association's view, the rule is a justified response to the irreproducibility crisis and reinforces the US government's long-standing commitment to base policy on the best available science.

What is good for «Science» ?

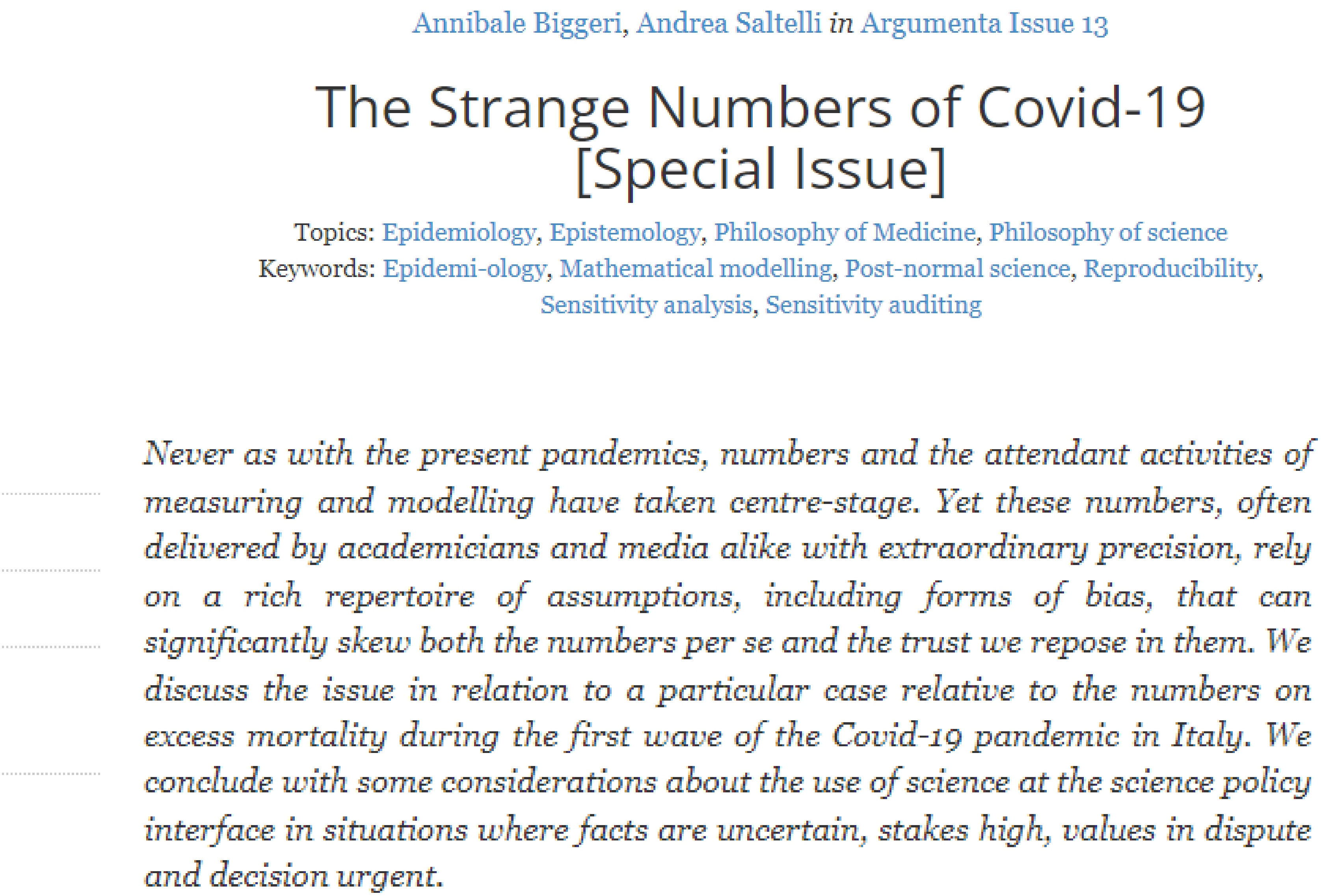
But the basic question is: “Is replicability good for science ?” As Bailey (2018) stated: “Different teams studying the same phenomena can often produce widely different results. (…) they may be also a sign of healthy scientific progress.”



Statistical Rituals: The Replication Delusion and How We Got There

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(2) 198–218

Gerd Gigerenzer
Harding Center for Risk Literacy, Max-Planck Institute for Human Development, Berlin, Germany



Annibale Biggeri, Andrea Saltelli *in* Argumenta Issue 13

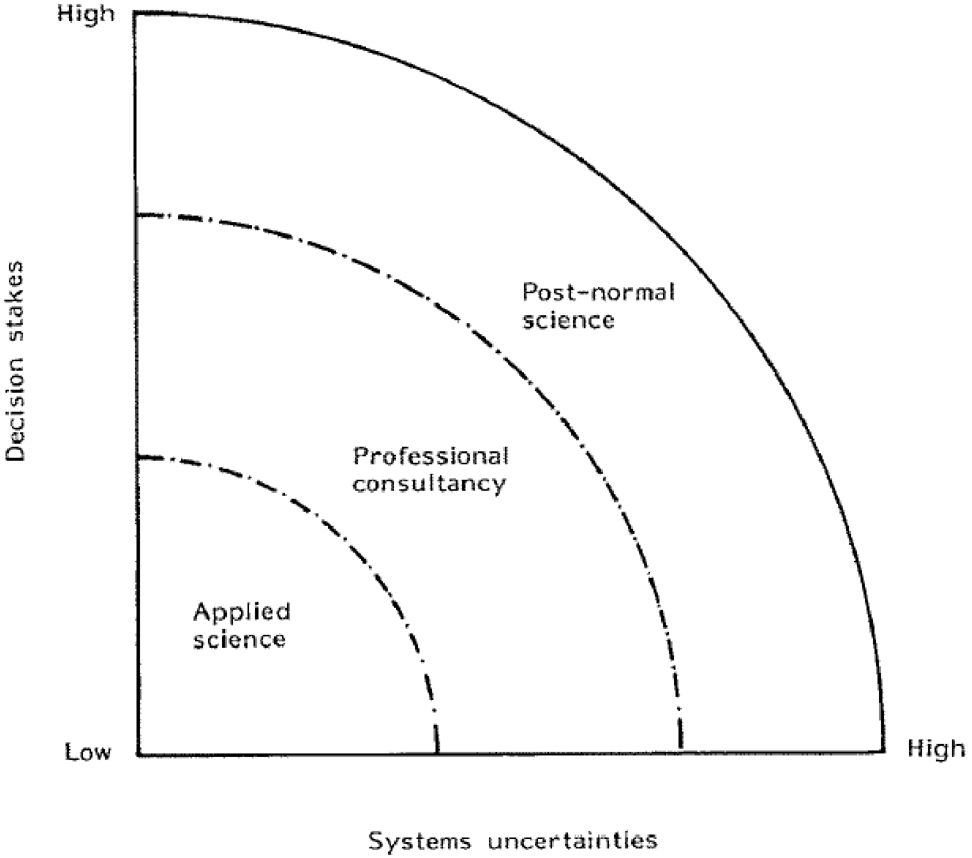
The Strange Numbers of Covid-19 [Special Issue]

Topics: Epidemiology, Epistemology, Philosophy of Medicine, Philosophy of science
Keywords: Epidemi-ology, Mathematical modelling, Post-normal science, Reproducibility, Sensitivity analysis, Sensitivity auditing

Never as with the present pandemics, numbers and the attendant activities of measuring and modelling have taken centre-stage. Yet these numbers, often delivered by academicians and media alike with extraordinary precision, rely on a rich repertoire of assumptions, including forms of bias, that can significantly skew both the numbers per se and the trust we repose in them. We discuss the issue in relation to a particular case relative to the numbers on excess mortality during the first wave of the Covid-19 pandemic in Italy. We conclude with some considerations about the use of science at the science policy interface in situations where facts are uncertain, stakes high, values in dispute and decision urgent.

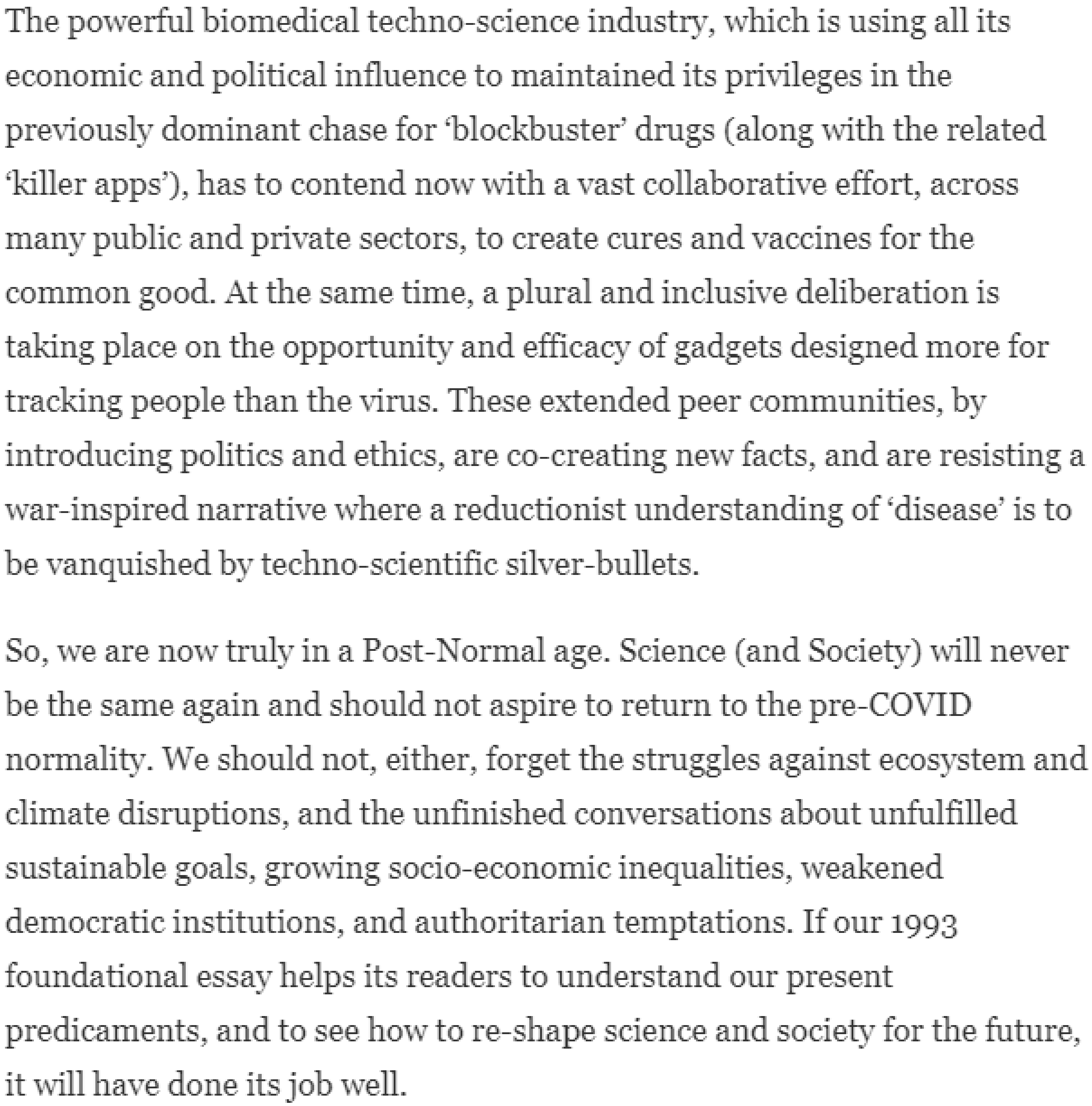
The powerful biomedical techno-science industry, which is using all its economic and political influence to maintained its privileges in the previously dominant chase for 'blockbuster' drugs (along with the related
'killer apps'), has to cc
many public and priv
common good. At the
taking place on the op
tracking people than t
introducing politics ar
war-inspired narrative
be vanquished by tech

So, we are now truly i
be the same again and
normality. We should
climate disruptions, a
sustainable goals, gro
democratic institution
foundational essay helps its readers to understand our present predicaments, and to see how to re-shape science and society for the future, it will have done its job well.



Funtowicz and Ravetz

FUTURES September 1993



The powerful biomedical techno-science industry, which is using all its economic and political influence to maintained its privileges in the previously dominant chase for 'blockbuster' drugs (along with the related 'killer apps'), has to contend now with a vast collaborative effort, across many public and private sectors, to create cures and vaccines for the common good. At the same time, a plural and inclusive deliberation is taking place on the opportunity and efficacy of gadgets designed more for tracking people than the virus. These extended peer communities, by introducing politics and ethics, are co-creating new facts, and are resisting a war-inspired narrative where a reductionist understanding of 'disease' is to be vanquished by techno-scientific silver-bullets.

So, we are now truly in a Post-Normal age. Science (and Society) will never be the same again and should not aspire to return to the pre-COVID normality. We should not, either, forget the struggles against ecosystem and climate disruptions, and the unfinished conversations about unfulfilled sustainable goals, growing socio-economic inequalities, weakened democratic institutions, and authoritarian temptations. If our 1993 foundational essay helps its readers to understand our present predicaments, and to see how to re-shape science and society for the future, it will have done its job well.

Re-print 2020 of
Funtowicz and Ravetz
FUTURES September 1993

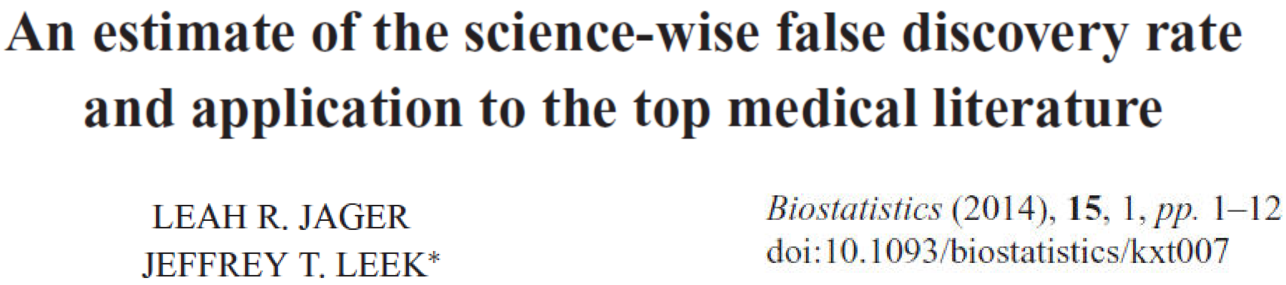
scheme of my talk

- The extent of the phenomenon
- A new discipline is born
- Data sharing
- Types of reproducibility
- Uncertainty of scientific measurements

- Daniel Sarewitz - “We have always been post-normal”

The extent of the phenomenon

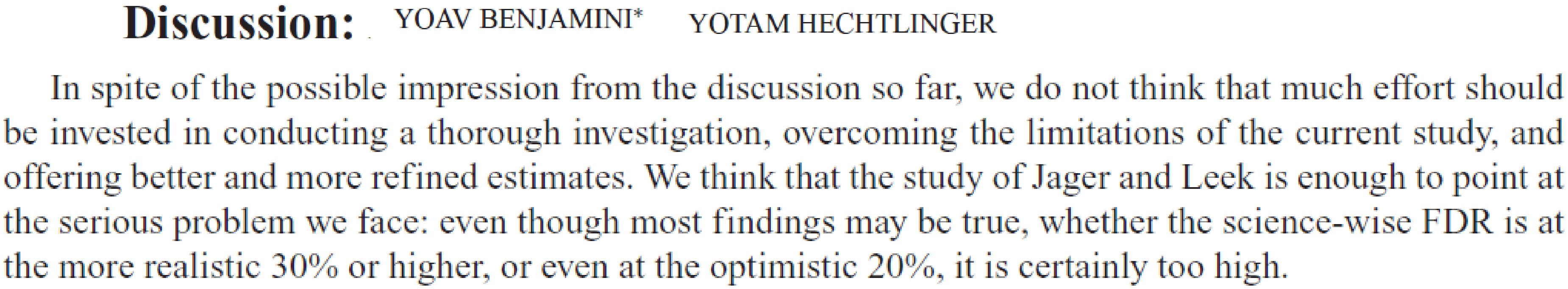
Scientist frame the problem as usual, starting with quantification:



An estimate of the science-wise false discovery rate and application to the top medical literature

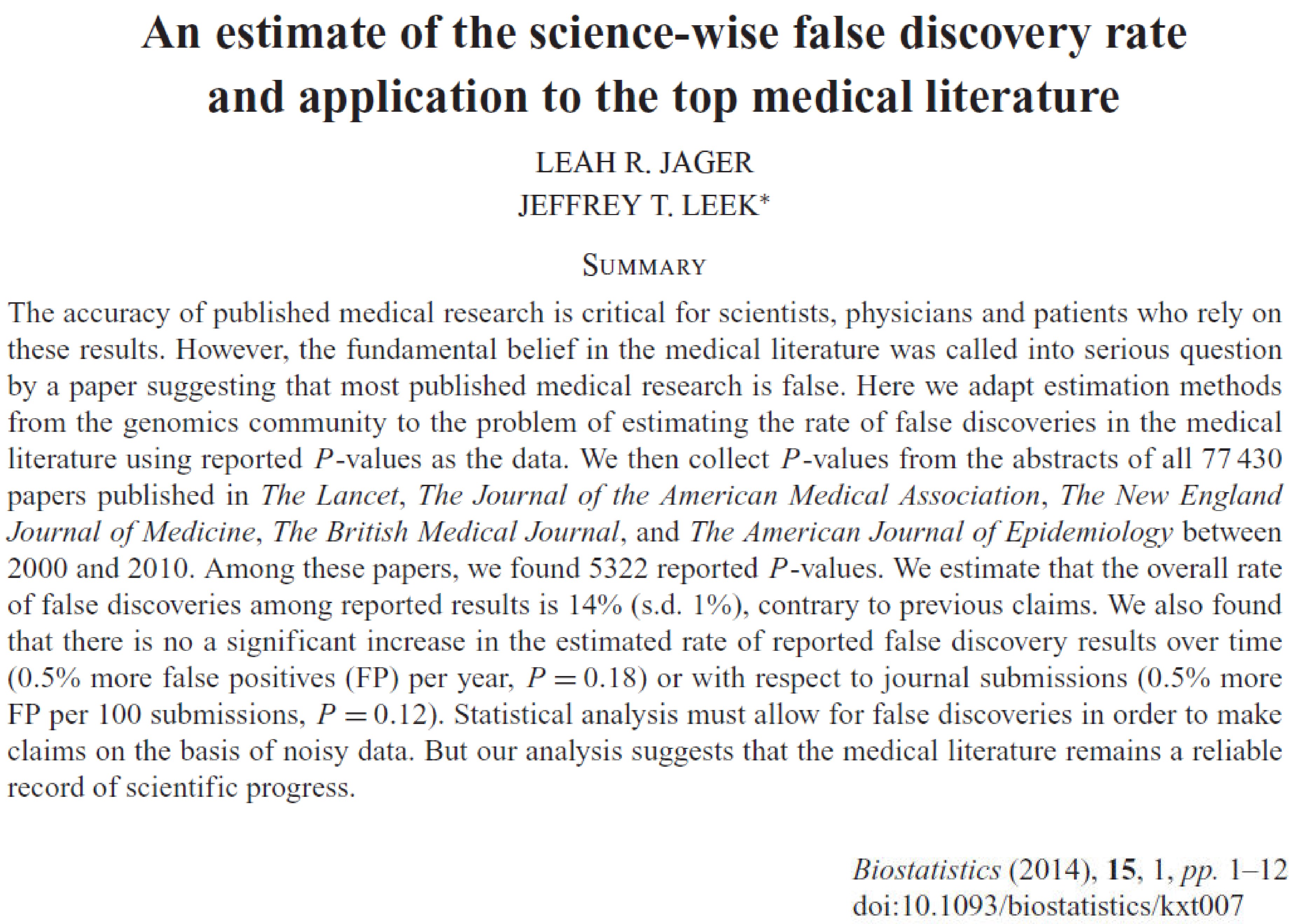
LEAH R. JAGER
JEFFREY T. LEEK*

Biostatistics (2014), **15**, 1, *pp.* 1–12
doi:10.1093/biostatistics/kxt007



Discussion: YOAV BENJAMINI* YOTAM HECHTLINGER

In spite of the possible impression from the discussion so far, we do not think that much effort should be invested in conducting a thorough investigation, overcoming the limitations of the current study, and offering better and more refined estimates. We think that the study of Jager and Leek is enough to point at the serious problem we face: even though most findings may be true, whether the science-wise FDR is at the more realistic 30% or higher, or even at the optimistic 20%, it is certainly too high.



An estimate of the science-wise false discovery rate and application to the top medical literature

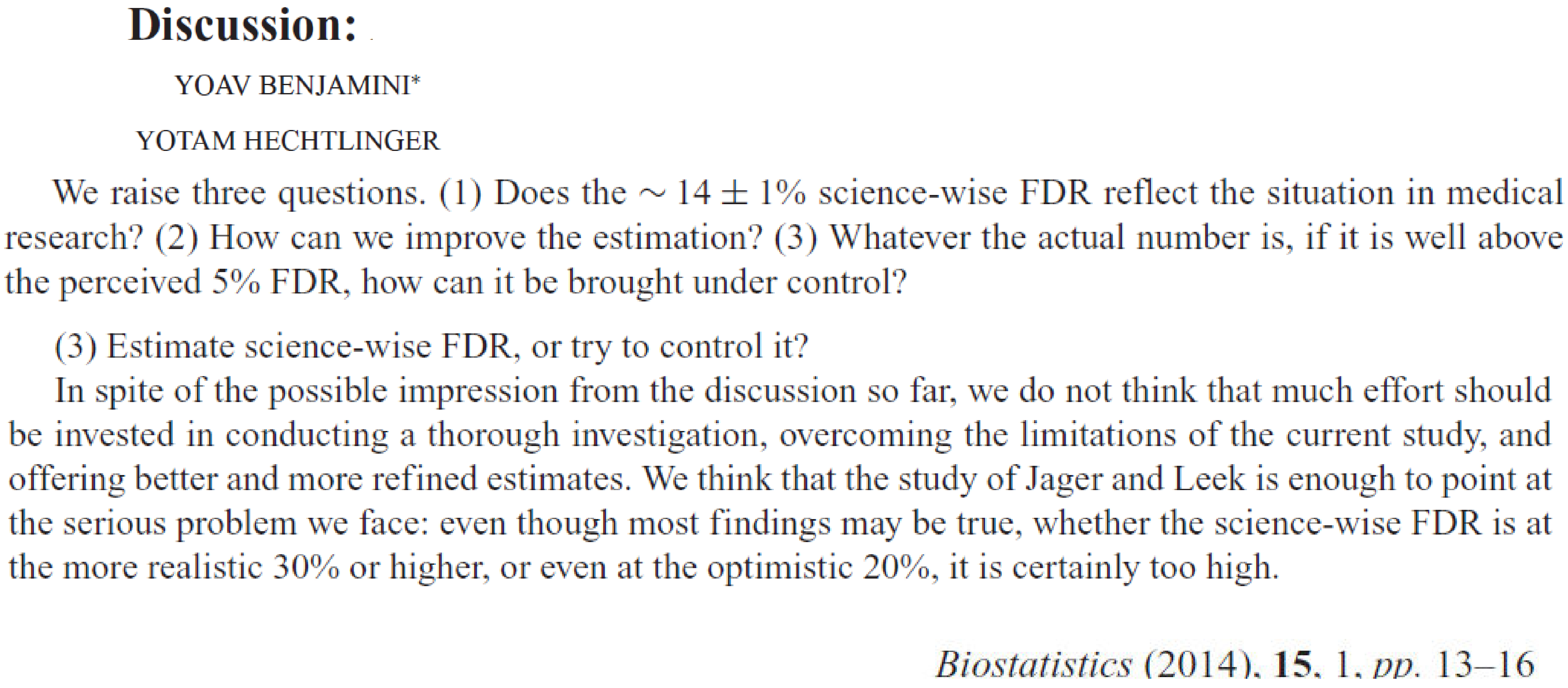
LEAH R. JAGER
JEFFREY T. LEEK*

SUMMARY

The accuracy of published medical research is critical for scientists, physicians and patients who rely on these results. However, the fundamental belief in the medical literature was called into serious question by a paper suggesting that most published medical research is false. Here we adapt estimation methods from the genomics community to the problem of estimating the rate of false discoveries in the medical literature using reported P-values as the data. We then collect P-values from the abstracts of all 77 430 papers published in *The Lancet*, *The Journal of the American Medical Association*, *The New England Journal of Medicine*, *The British Medical Journal*, and *The American Journal of Epidemiology* between 2000 and 2010. Among these papers, we found 5322 reported P-values. We estimate that the overall rate of false discoveries among reported results is 14% (s.d. 1%), contrary to previous claims. We also found that there is no a significant increase in the estimated rate of reported false discovery results over time (0.5% more false positives (FP) per year, $P = 0.18$) or with respect to journal submissions (0.5% more FP per 100 submissions, $P = 0.12$). Statistical analysis must allow for false discoveries in order to make claims on the basis of noisy data. But our analysis suggests that the medical literature remains a reliable record of scientific progress.

Biostatistics (2014), **15**, 1, *pp.* 1–12
doi:10.1093/biostatistics/kxt007

A better approach would be control of False Coverage Rate (Yekutieli e Benjamini 2005) in order to adjust for selective reporting.



Discussion:

YOAV BENJAMINI*

YOTAM HECHTLINGER

We raise three questions. (1) Does the $\sim 14 \pm 1\%$ science-wise FDR reflect the situation in medical research? (2) How can we improve the estimation? (3) Whatever the actual number is, if it is well above the perceived 5% FDR, how can it be brought under control?

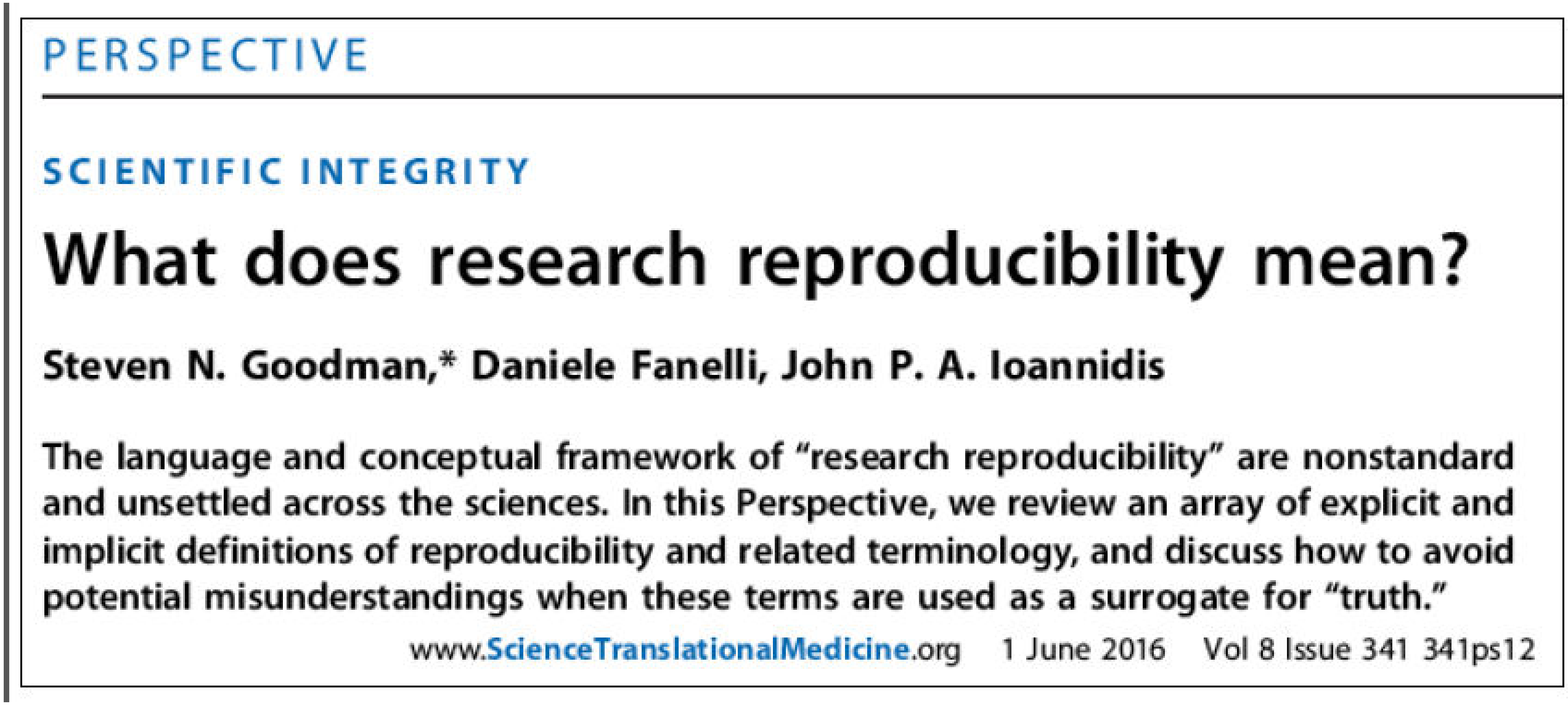
(3) Estimate science-wise FDR, or try to control it?

In spite of the possible impression from the discussion so far, we do not think that much effort should be invested in conducting a thorough investigation, overcoming the limitations of the current study, and offering better and more refined estimates. We think that the study of Jager and Leek is enough to point at the serious problem we face: even though most findings may be true, whether the science-wise FDR is at the more realistic 30% or higher, or even at the optimistic 20%, it is certainly too high.

Biostatistics (2014), **15**, 1, *pp.* 13–16

A new discipline is born

METRICS
META·RESEARCH INNOVATION
CENTER AT STANFORD



PERSPECTIVE

SCIENTIFIC INTEGRITY

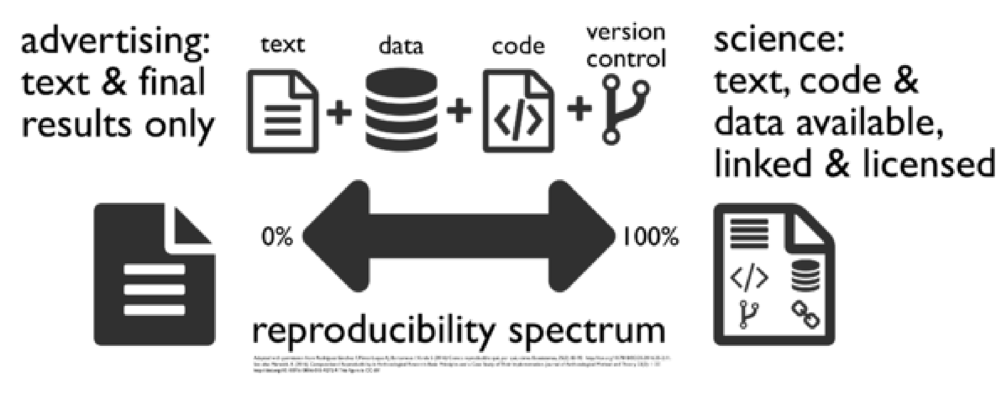
What does research reproducibility mean?

Steven N. Goodman,* Daniele Fanelli, John P. A. Ioannidis

The language and conceptual framework of "research reproducibility" are nonstandard and unsettled across the sciences. In this Perspective, we review an array of explicit and implicit definitions of reproducibility and related terminology, and discuss how to avoid potential misunderstandings when these terms are used as a surrogate for "truth."

www.ScienceTranslationalMedicine.org 1 June 2016 Vol 8 Issue 341 341ps12

Data sharing
Open Science
Intellectual property


advertising: text & final results only
text
data
code
version control
science: text, code & data available, linked & licensed
0%
100%
reproducibility spectrum

The reproducible research spectrum. Reproducibility is not a binary quality but a spectrum. Scientific articles that contain only the final text, results and figures (e.g., in a single pdf document) are advertising a finding, and these are the least reproducible -- it is often impossible to reconstruct the whole analytical process from data to results. Publication of the data and/or code used for the analysis greatly improves reproducibility. Similarly, using a version control system (such as git) permits navigating through the complete history of the project. Finally, the most reproducible, and thus scientific, studies are those using dynamic reports (e.g., Rmarkdown notebooks) that integrate text, code and data into an executable environment.

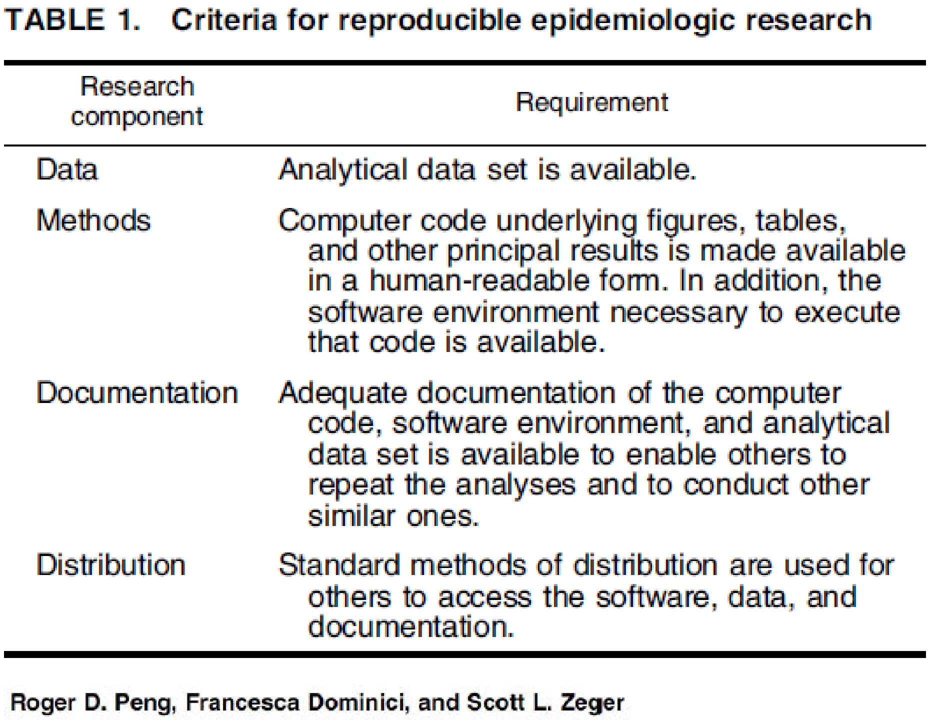


TABLE 1. Criteria for reproducible epidemiologic research

Research component	Requirement
Data	Analytical data set is available.
Methods	Computer code underlying figures, tables, and other principal results is made available in a human-readable form. In addition, the software environment necessary to execute that code is available.
Documentation	Adequate documentation of the computer code, software environment, and analytical data set is available to enable others to repeat the analyses and to conduct other similar ones.
Distribution	Standard methods of distribution are used for others to access the software, data, and documentation.

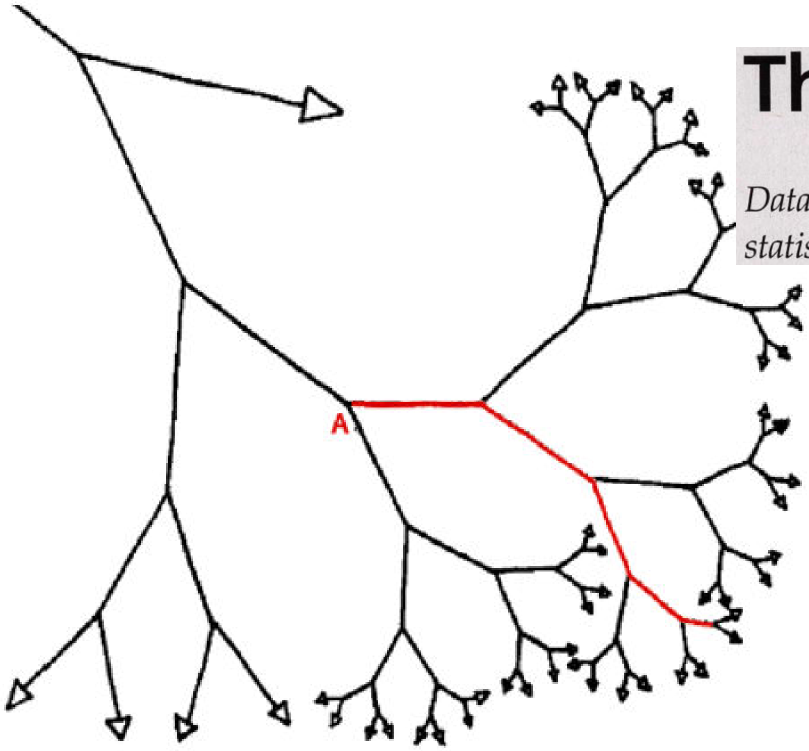
Roger D. Peng, Francesca Dominici, and Scott L. Zeger

Dynamic reporting:

R markdown
Stata MarkDoc
Weaver
Knitr

…….

Re-analysis of the same datset ?



Gelman & Loken (2014)

The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

JORGE LUIS BORGES

EL JARDIN DE SENDEROS QUE SE BIFURCAN

SUR

BUENOS AIRES

CORRESPONDENCE

Re: Glyphosate Use and Cancer Incidence in the Agricultural Health Study

Lianne Sheppard, Rachel M. Shaffer

See the Notes section for the full list of authors' affiliations.
Correspondence to: Lianne Sheppard, PhD, Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195-7232 (e-mail: sheppard@uw.edu).

Exposure to glyphosate, a broad-spectrum herbicide, and its consequent health impacts are critically important to understand. Its use and potential to enter the food supply have increased dramatically worldwide over recent decades (1). The Agricultural Health Study (AHS) is a crucial piece of evidence because there are no other large cohort studies of the potential carcinogenic effects of glyphosate. Thus, the recent AHS results (2), adding 11 years of follow-up to the previously reported AHS results (3), have huge potential to improve our understanding of glyphosate toxicity—which in turn informs national and international evaluations that influence policy. We wish to describe a feature of the study analyses that most likely attenuated the effect estimates towards the null. The exposure modeling introduced noise into the exposure estimates because it used a multiple imputation procedure that did not consider the study outcomes (4).

Exposure assessment in the AHS was complete at baseline. The investigators faced a huge challenge because 20 968 individuals, 37% of AHS participants, failed to complete the follow-up questionnaire. The authors used a well-respected approach to fill in missing data with multiple draws of the distribution of the missing exposures conditional on information available from baseline, including demographics, farm characteristics, pesticide use history, and existing medical conditions (4). They called their procedure multiple imputation, because it has the appearance of the multiple imputation approach described by Rubin (5). However, the AHS multiple imputation did not consider any of the health outcomes analyzed by Andreotti et al. (2), including non-Hodgkin lymphoma and multiple myeloma. As was elegantly stated in his review, "Regression with missing X's," Little noted that when realizations of the distribution of missing exposures (X_1 in his notation) given other covariate data ($X2, \ldots, X_p$) are used to estimate regression coefficients, failure to condition on the health outcome (Y) leads to bias. The direction of the bias is attenuation towards no increased risk. Little explained that for an exposure X_1, "… then the regression coefficient of X_1 is attenuated, because the noise added to the conditional means doesn't account for the partial correlation of X_1 and Y given $X_2, \ldots, X_p$." (p. 1235) (6). Gryparis et al. (7) name this misguided approach to multiple imputation "exposure simulation."

We do not know the size of the exposure model residual in the AHS or the magnitude of the resulting bias. However, because the phase II nonrespondent group was large (37%) and the phase II respondent group reported a high prevalence of glyphosate use (52%), there is reason to suspect that the consequence of using this imputation procedure is to meaningfully attenuate the cancer risk estimates. Unfortunately, it is unlikely that the conclusion of Heltshe et al. (4)—"This multiple imputation will allow for bias reduction and improved efficiency in future analyses of the AHS"—is correct. We encourage the AHS investigators to refine their approach and improve our ability to understand the true impacts of pesticide exposures, which—particularly for glyphosate—could have tangible consequences for public health policy.

Funding

RMS is supported by NIEHS T32ES015459.

Notes

Affiliations of authors: Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, WA (LS, RMS); Department of Biostatistics, University of Washington, Seattle, WA (LS).

Dr Sheppard was an ad hoc member of the EPA Federal Insecticide, Rodenticide, and Fungicide Act Scientific Advisory Panel for Evaluation of the Carcinogenic Potential of Glyphosate in 2016–2017.

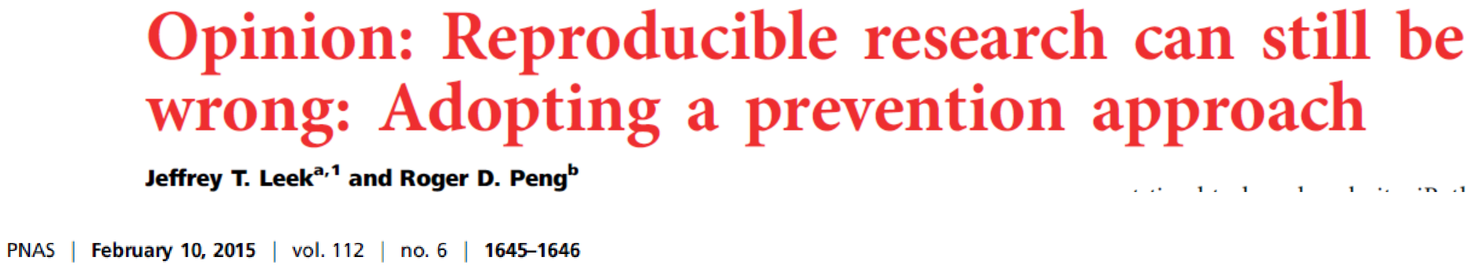
Types of reproducibility

Goodman defines different types of reproducibility:

- Methodological
- of Results
- Scientific

- Methodological: it was immediately clear that this is a weak argument and doubt strategies may be easily played (see EPA's statement).

Again an "internal solution" is discussed:



Opinion: Reproducible research can still be wrong: Adopting a prevention approach

Jeffrey T. Leek[a,1] and Roger D. Peng[b]

PNAS | February 10, 2015 | vol. 112 | no. 6 | 1645–1646

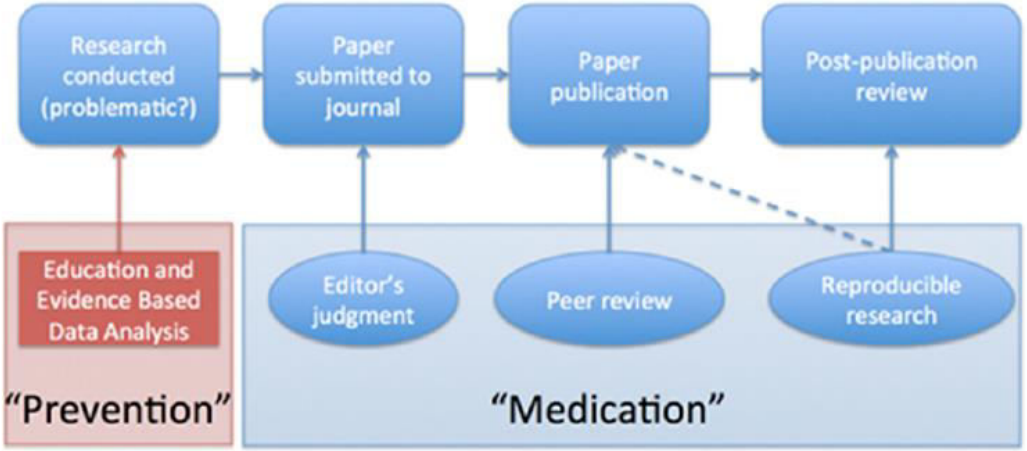
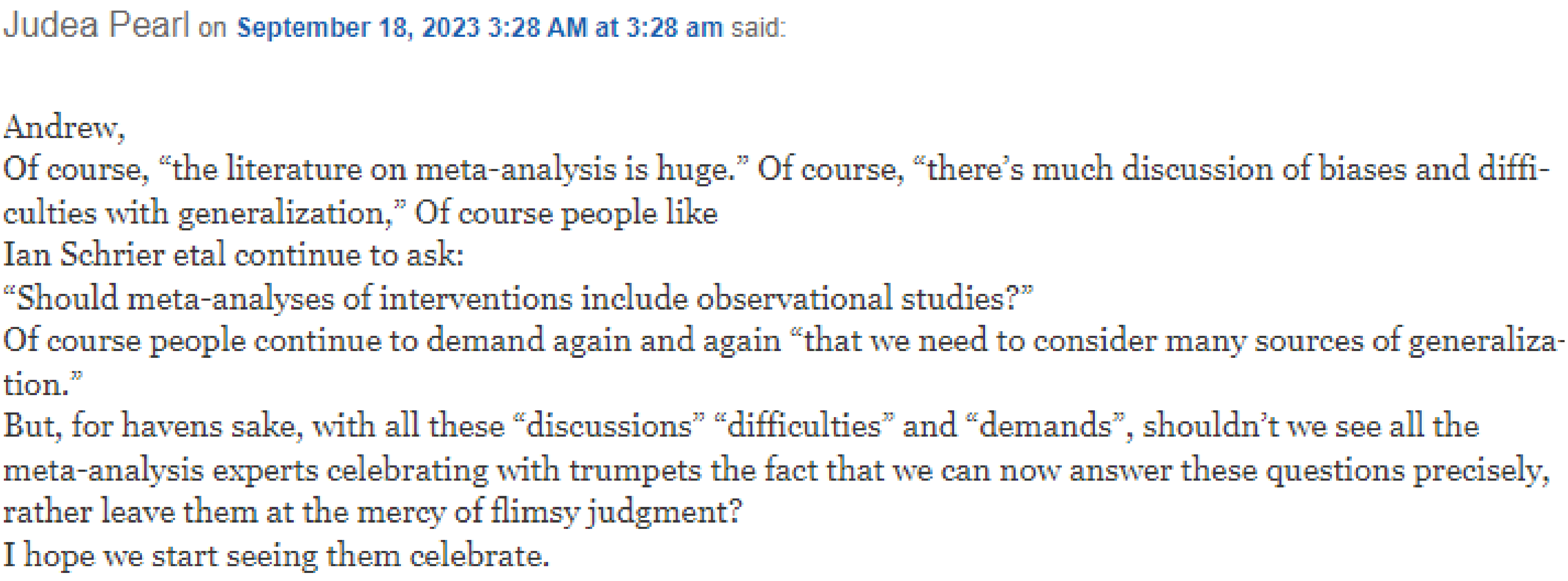


Fig. 1. Peer review and editor evaluation help treat poor data analysis. Education and evidence-based data analysis can be thought of as preventative measures.

- of Results: this applies to all proposed methods of synthesis (metanalysis, systematic reviews and so on).



Judea Pearl on September 18, 2023 3:28 AM at 3:28 am said:

Andrew,
Of course, “the literature on meta-analysis is huge.” Of course, “there’s much discussion of biases and difficulties with generalization,” Of course people like
Ian Schrier etal continue to ask:
“Should meta-analyses of interventions include observational studies?”
Of course people continue to demand again and again “that we need to consider many sources of generalization.”
But, for havens sake, with all these “discussions” “difficulties” and “demands”, shouldn’t we see all the meta-analysis experts celebrating with trumpets the fact that we can now answer these questions precisely, rather leave them at the mercy of flimsy judgment?
I hope we start seeing them celebrate.

Let’s consider the case of two studies.

		Study 2	
		$\theta=1$	**$\theta>1$**
Study 1	**$\theta=1$**	**A**	**B**
	$\theta>1$	**C**	**D**

Null hypothesis is usually rejected when at least one study does reject it (Ho corresponds to the joint event A in the table)

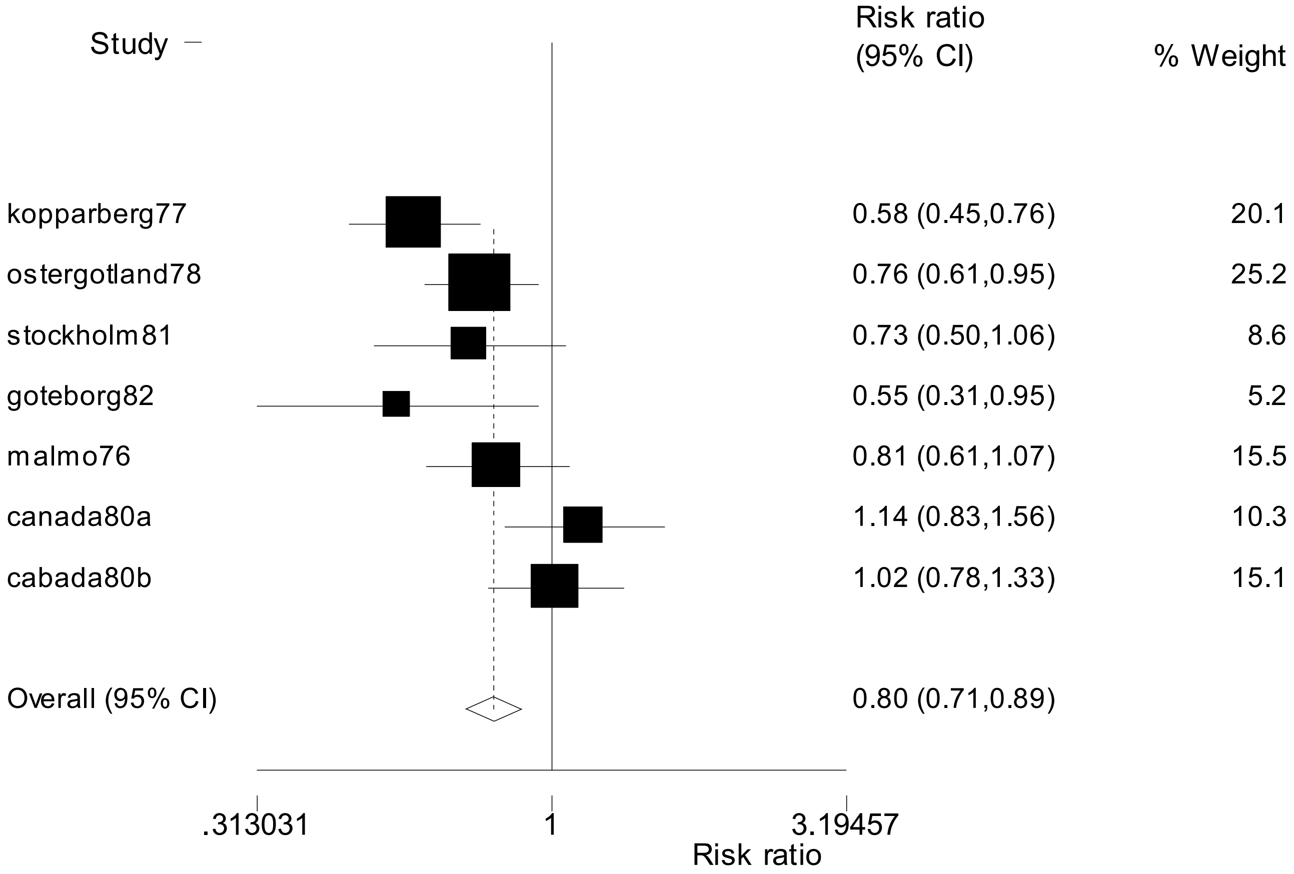
The reproducibility null hyptohesis is rejected when all studies reject $\theta=1$ (Ho correspond to the event $A\cup B\cup C$ in the table).

Benjamini, Heller and Yekutieli (2009) defined the **r-value**, in this simple case r=max(p1,p2), the p-values of the two studies.

We show the contradictions using a metanalysis on breast cancer screening.

Olsen and Goetzche conducted a metanalysis on breast cancer screening efficacy (Lancet 2001). Here we consider the outcome “Breast cancer specific mortality – 13 years follow-up” (table 9).

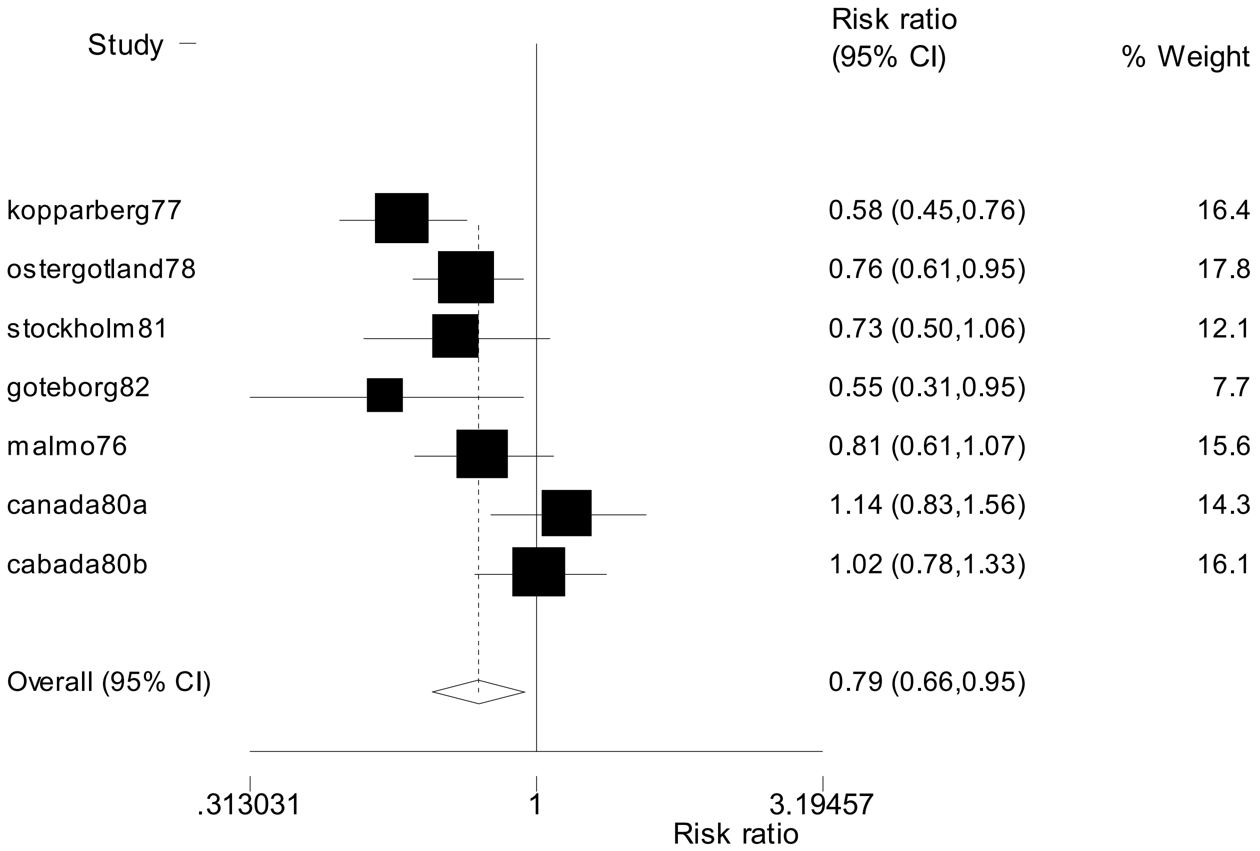
```
           Study |       RR    [95% CI]       % Weight
-----------------+------------------------------------
malmo76          |  0.81     0.61   1.07      15.4608
canada80a        |  1.148    0.83   1.56      10.3287
cabada80b        |  1.02     0.78   1.33      15.0698
kopparberg77     |  0.58     0.45   0.76      20.141
ostergotland78   |  0.76     0.61   0.95      25.1743
stockholm81      |  0.72     0.50   1.06      8.63844
goteborg82       |  0.55     0.31   0.95      5.18692
-----------------+------------------------------------
  M-H pooled RR  |  0.80     0.71   0.89
-----------------+------------------------------------
  Heterogeneity chi-squared =  15.89 (d.f. = 6) p = 0.014
  Test of RR=1 : z= 4.07 p = 0.000
```



Study	Risk ratio (95% CI)	% Weight
kopparberg77	0.58 (0.45,0.76)	20.1
ostergotland78	0.76 (0.61,0.95)	25.2
stockholm81	0.73 (0.50,1.06)	8.6
goteborg82	0.55 (0.31,0.95)	5.2
malmo76	0.81 (0.61,1.07)	15.5
canada80a	1.14 (0.83,1.56)	10.3
cabada80b	1.02 (0.78,1.33)	15.1
Overall (95% CI)	0.80 (0.71,0.89)	

Are the study results homogeneous ? Random effect model results:

```
           Study |       RR    [95% CI]       % Weight
-----------------+---------------------------------------
malmo76          |  0.81     0.61   1.07      15.5547
canada80a        |  1.148    0.83   1.56      14.2565
cabada80b        |  1.02     0.78   1.33      16.0716
kopparberg77     |  0.58     0.45   0.76      16.4466
ostergotland78   |  0.76     0.61   0.95      17.8396
stockholm81      |  0.72     0.50   1.06      12.1202
goteborg82       |  0.55     0.31   0.95      7.71074
-----------------+---------------------------------------
 D+L pooled RR   |  0.79     0.66   0.95
-----------------+---------------------------------------
  Heterogeneity chi-squared =  15.89 (d.f. = 6) p = 0.014
  Estimate of between-study variance Tau-squared = 0.0381
  Test of RR=1 : z= 2.46 p = 0.014
```


Study
Risk ratio
(95% CI)
% Weight
kopparberg77
0.58 (0.45,0.76)
16.4
ostergotland78
0.76 (0.61,0.95)
17.8
stockholm81
0.73 (0.50,1.06)
12.1
goteborg82
0.55 (0.31,0.95)
7.7
malmo76
0.81 (0.61,1.07)
15.6
canada80a
1.14 (0.83,1.56)
14.3
cabada80b
1.02 (0.78,1.33)
16.1
Overall (95% CI)
0.79 (0.66,0.95)
.313031
1
3.19457
Risk ratio

Heterogeneity does not necessarily implies lack of replicability.

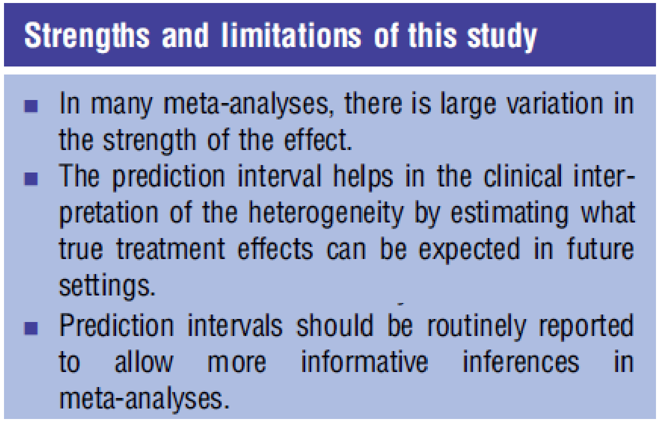
A consequence of this misunterstanding is the use of predictive intervals

$$\mathrm{SD}_{\mathrm{PI}} = \surd(\tau^2+\mathrm{SE}^2)$$

95% PrI 0.54; 1.16

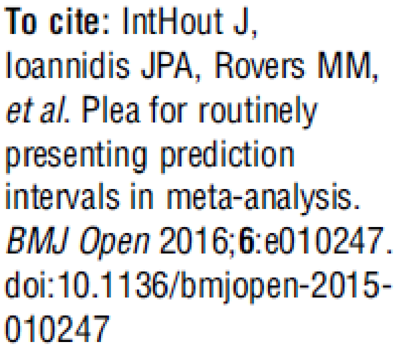
95% RE-CI 0.60; 0.95

95% FE-CI 0.71; 0.89



Strengths and limitations of this study

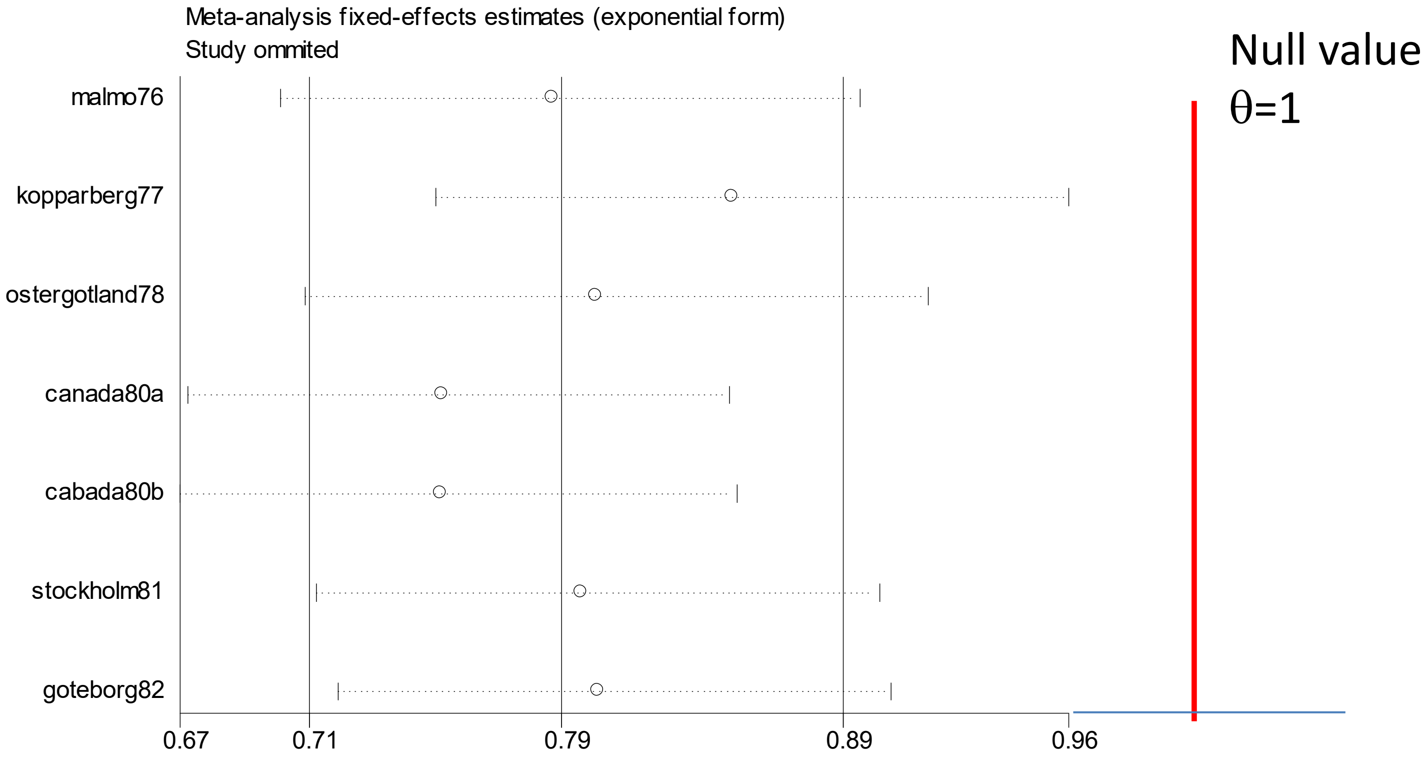
- In many meta-analyses, there is large variation in the strength of the effect.
- The prediction interval helps in the clinical interpretation of the heterogeneity by estimating what true treatment effects can be expected in future settings.
- Prediction intervals should be routinely reported to allow more informative inferences in meta-analyses.



To cite: IntHout J, Ioannidis JPA, Rovers MM, *et al.* Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open* 2016;**6**:e010247. doi:10.1136/bmjopen-2015-010247

r-value

```
---------------------------------------------------------
 Study ommited      |   e^coef.[95%  Conf. Interval]
-------------------+-------------------------------------
 malmo76            |   0.79      0.70    0.89
 kopparberg77       |   0.85      0.75    0.96
 ostergotland78     |   0.80      0.71    0.91
 canada80a          |   0.75      0.67    0.85
 cabada80b          |   0.75      0.67    0.85
 stockholm81        |   0.80      0.71    0.90
 goteborg82         |   0.81      0.72    0.90
-------------------+-------------------------------------
```



r-value corresponds to the p-value omitting study n 2

IARC-NCI workshop on an epidemiological toolkit to assess biases in human cancer studies for hazard identification: beyond the algorithm

Mary K Schubauer-Berigan,[1] David B Richardson,[2] Matthew P Fox,[3] Lin Fritschi,[4] Irina Guseva Canu,[5] Neil Pearce,[6] Leslie Stayner,[1] Amy Berrington de Gonzalez[7,8]

Occup Env Med 2023

A risk of bias instrument for non-randomized studies of exposures: A users' guide to its application in the context of GRADE

Rebecca L. Morgan[a], Kristina A. Thayer[b], Nancy Santesso[a], Alison C. Holloway[c], Robyn Blain[d], Sorina E. Eftim[d], Alexandra E. Goldstone[d], Pam Ross[d], Mohammed Ansari[e], Elie A Akl[a,f], Tommaso Filippini[g], Anna Hansell[h,i,j], Joerg J. Meerpohl[k], Reem A. Mustafa[a,l], Jos Verbeek[m], Marco Vinceti[g,n], Paul Whaley[o], Holger J. Schünemann[a,p,*], GRADE Working Group

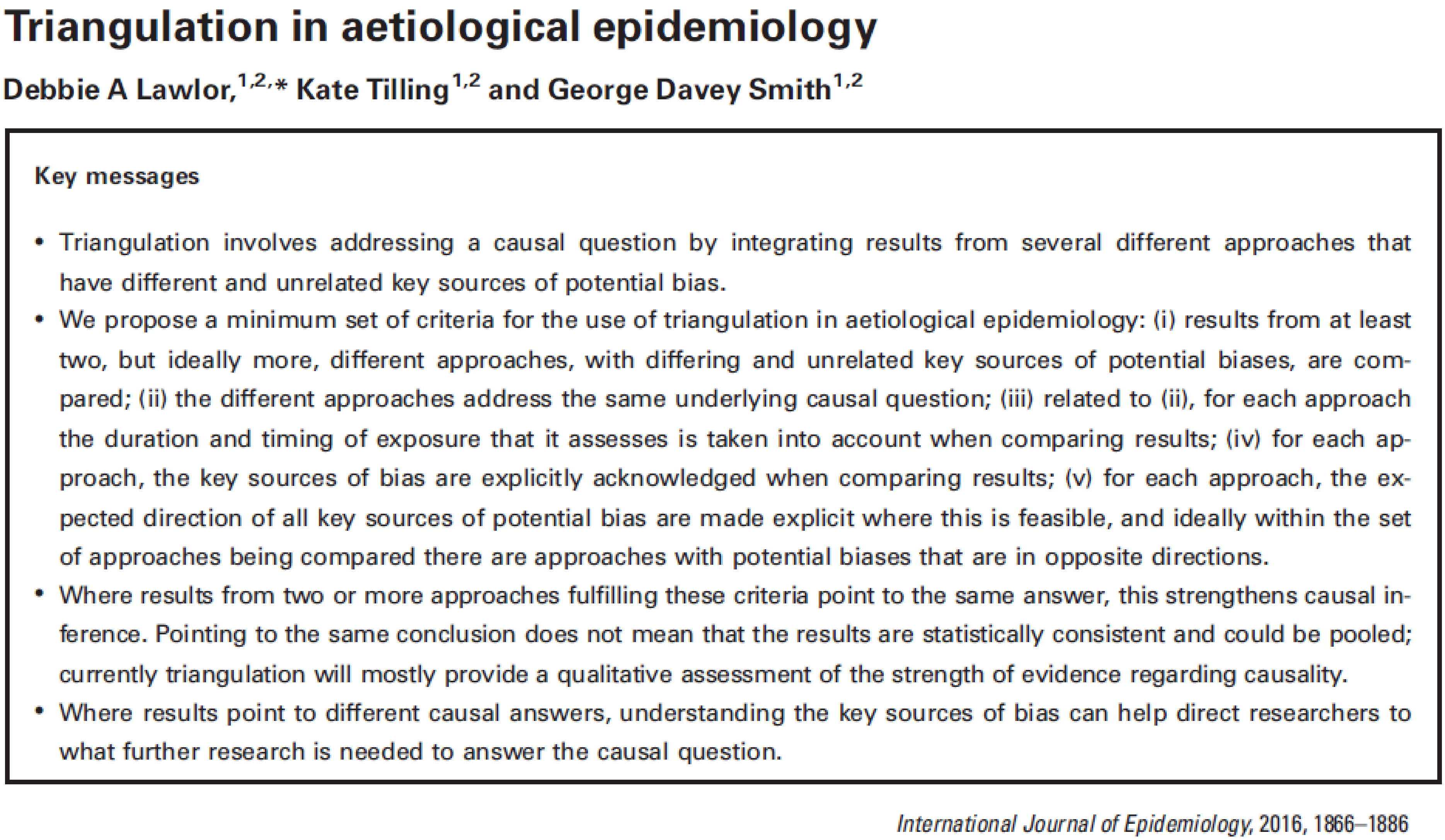
Env Int 2019

Risk of Bias Assessments and Evidence Syntheses for Observational Epidemiologic Studies of Environmental and Occupational Exposures: Strengths and Limitations

Kyle Steenland,[1] M.K. Schubauer-Berigan,[2] R. Vermeulen,[3] R.M. Lunn,[4] K. Straif,[5,6] S. Zahm,[7] P. Stewart,[8] W.D. Arroyave,[9] S.S. Mehta,[4] and N. Pearce[10]

Environmental Health Perspectives 095002-8 128(9) September 2020

- Scientific: Goodman (2018) cites William Whewell (1794-1866) and the concept of Consilience



Triangulation in aetiological epidemiology

Debbie A Lawlor,[1,2,*] Kate Tilling[1,2] and George Davey Smith[1,2]

Key messages

- Triangulation involves addressing a causal question by integrating results from several different approaches that have different and unrelated key sources of potential bias.
- We propose a minimum set of criteria for the use of triangulation in aetiological epidemiology: (i) results from at least two, but ideally more, different approaches, with differing and unrelated key sources of potential biases, are compared; (ii) the different approaches address the same underlying causal question; (iii) related to (ii), for each approach the duration and timing of exposure that it assesses is taken into account when comparing results; (iv) for each approach, the key sources of bias are explicitly acknowledged when comparing results; (v) for each approach, the expected direction of all key sources of potential bias are made explicit where this is feasible, and ideally within the set of approaches being compared there are approaches with potential biases that are in opposite directions.
- Where results from two or more approaches fulfilling these criteria point to the same answer, this strengthens causal inference. Pointing to the same conclusion does not mean that the results are statistically consistent and could be pooled; currently triangulation will mostly provide a qualitative assessment of the strength of evidence regarding causality.
- Where results point to different causal answers, understanding the key sources of bias can help direct researchers to what further research is needed to answer the causal question.

International Journal of Epidemiology, 2016, 1866–1886

- Definition
- Framing
- Relationship with causal inference
- How it works

Definition

- The term came from philosophy of science, the original concept was “consilience”

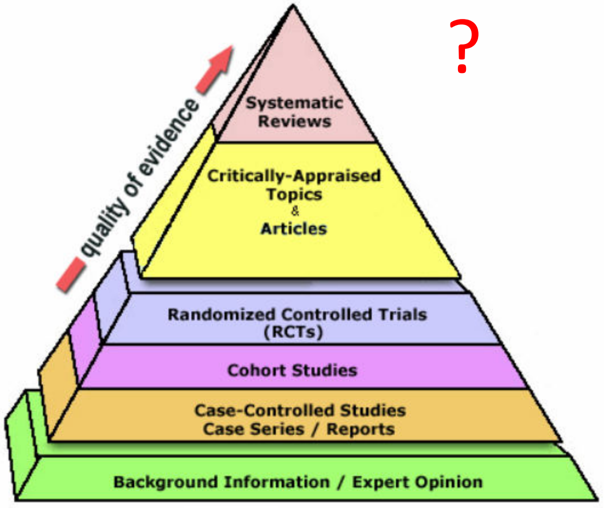
Goodman (HEI 2018) recalled William Whewell (1794-1866) who introduced the concept of *Consilience* : the method to prove a theory or estimating a given parameter by means of multiple measurements taken in several different experimental modalities chosen in a way to assure that they do not share the same sources of bias.

Definition: a simple exercise

Let us compare results from a case-control study and a cohort study.

1- list major menaces to validity in a case-control study

2- list major menaces to validity in a cohort study


?
quality of evidence
Systematic Reviews
Critically-Appraised Topics & Articles
Randomized Controlled Trials (RCTs)
Cohort Studies
Case-Controlled Studies Case Series / Reports
Background Information / Expert Opinion

A hierarchy of study types.

Definition: a simple exercise

Let us compare results from a case-control study and a cohort study.

1- list major advantages of a case-control study

2- list major advantages in a cohort study

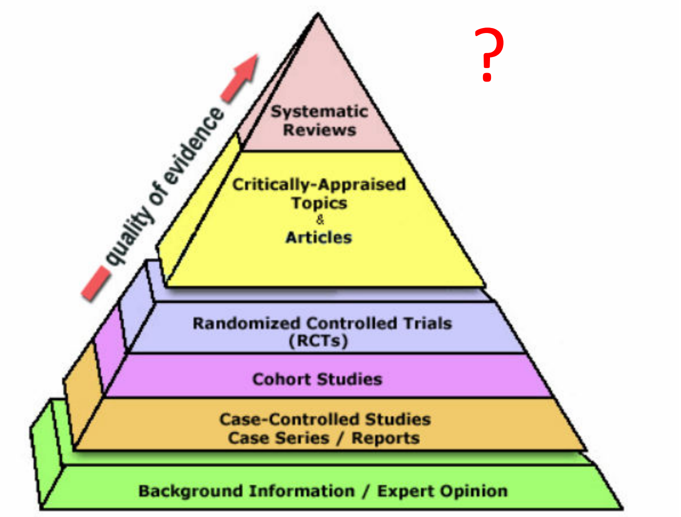

?
quality of evidence
Systematic Reviews
Critically-Appraised Topics & Articles
Randomized Controlled Trials (RCTs)
Cohort Studies
Case-Controlled Studies Case Series / Reports
Background Information / Expert Opinion

Figure 4. A hierarchy of study types.

Definition: a simple exercise

Let us compare results from a case-control study and a cohort study.

1- list major menaces to validity in a case-control study

Recall bias, Reverse causation, Exposure misclassification

2- list major menaces to validity in a cohort study

Confounding, Outcome misclassification

Definition: a simple exercise

Let us compare results from a case-control study and a cohort study.

1- list major advantages of a case-control study

Outcome definition, Confounders definition

2- list major advantages in a cohort study

Exposure definition, time-dependent confounding, cause preceding the effect

Definition: a simple exercise

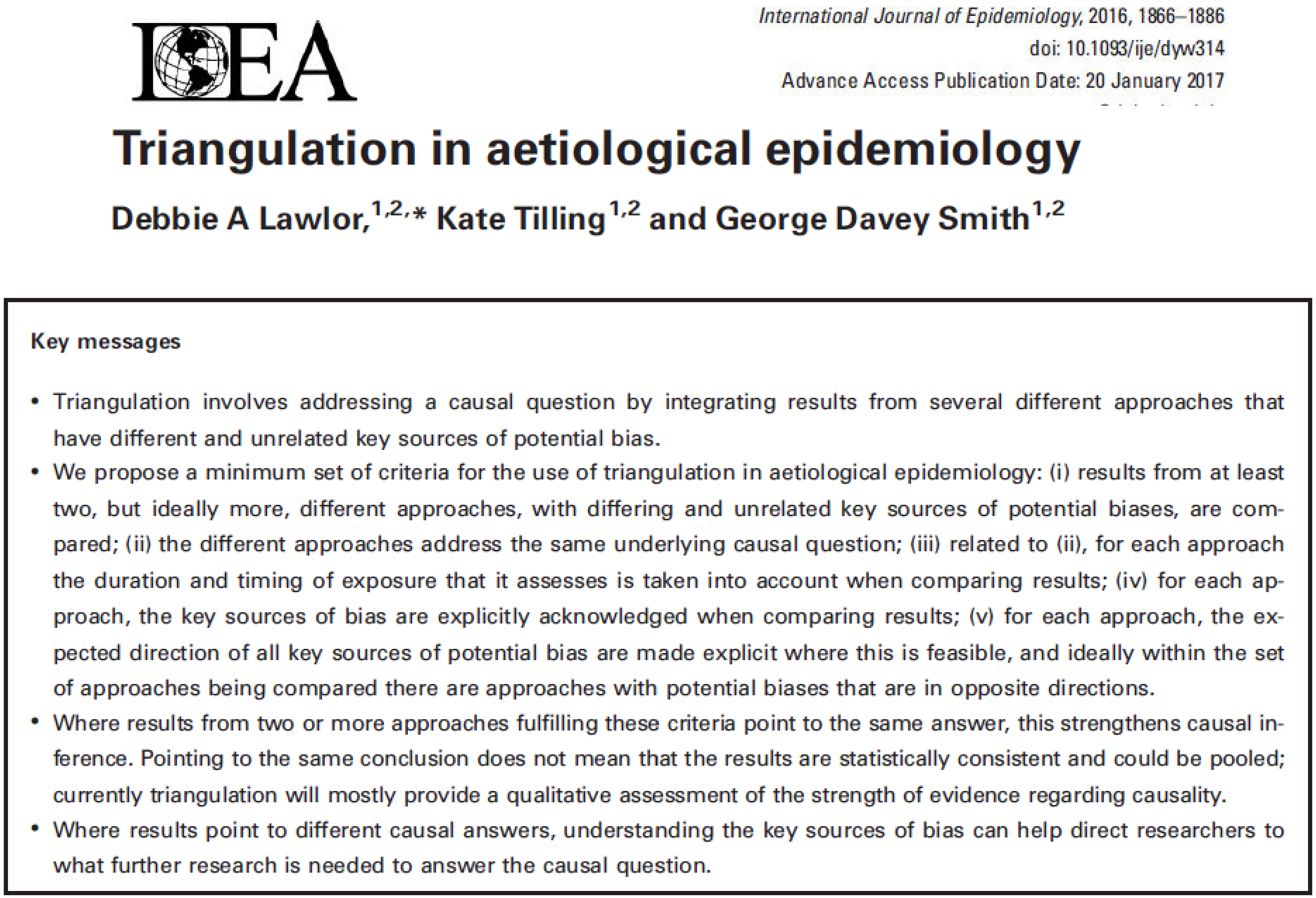
Design issues of a case-control study and a cohort study.

1- case-control study

One disease, multiple potential exposures, RR overestimated

2- cohort study

One exposure, multiple diseases, RR underestimated



International Journal of Epidemiology, 2016, 1866–1886
doi: 10.1093/ije/dyw314
Advance Access Publication Date: 20 January 2017

Triangulation in aetiological epidemiology

Debbie A Lawlor,[1,2,*] Kate Tilling[1,2] and George Davey Smith[1,2]

Key messages

- Triangulation involves addressing a causal question by integrating results from several different approaches that have different and unrelated key sources of potential bias.
- We propose a minimum set of criteria for the use of triangulation in aetiological epidemiology: (i) results from at least two, but ideally more, different approaches, with differing and unrelated key sources of potential biases, are compared; (ii) the different approaches address the same underlying causal question; (iii) related to (ii), for each approach the duration and timing of exposure that it assesses is taken into account when comparing results; (iv) for each approach, the key sources of bias are explicitly acknowledged when comparing results; (v) for each approach, the expected direction of all key sources of potential bias are made explicit where this is feasible, and ideally within the set of approaches being compared there are approaches with potential biases that are in opposite directions.
- Where results from two or more approaches fulfilling these criteria point to the same answer, this strengthens causal inference. Pointing to the same conclusion does not mean that the results are statistically consistent and could be pooled; currently triangulation will mostly provide a qualitative assessment of the strength of evidence regarding causality.
- Where results point to different causal answers, understanding the key sources of bias can help direct researchers to what further research is needed to answer the causal question.

Framing

- Extending meta-analysis to include different study designs ?
- Improving synthesis of evidence wrt risk-of-bias assessment ?
- Strengthening causal inference ?
- Supporting scientific reproducibility ?

- Definition: consilience, i.e. coherence of different approaches
- Framing: scientific reproducibility
- Relationship with causal inference
- How it works

Relationship with causal inference

As stated by Hammerton and Munafò

“One reason to include design-based approaches is that these may be less likely to suffer from similar sources and directions of bias compared with statistical approaches, particularly when these are conducted within the same data set (Lawlor, Tilling, & DaveySmith, 2016).

Ideally, we would identify different sources of evidence that we could apply to a research question and understand the likely sources and directions of bias operating within each so that we could ensure that these are different.

This means that triangulation should be a prospective approach, rather than simply selecting sources of evidence that support a particular conclusion post hoc”

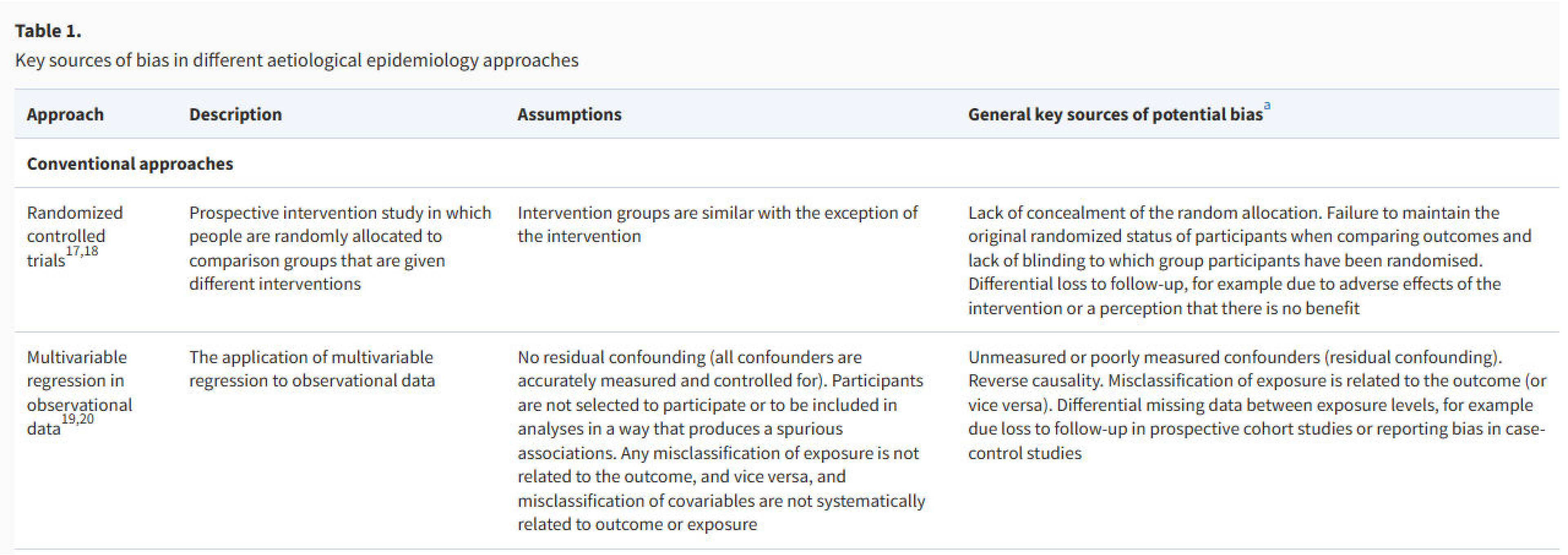
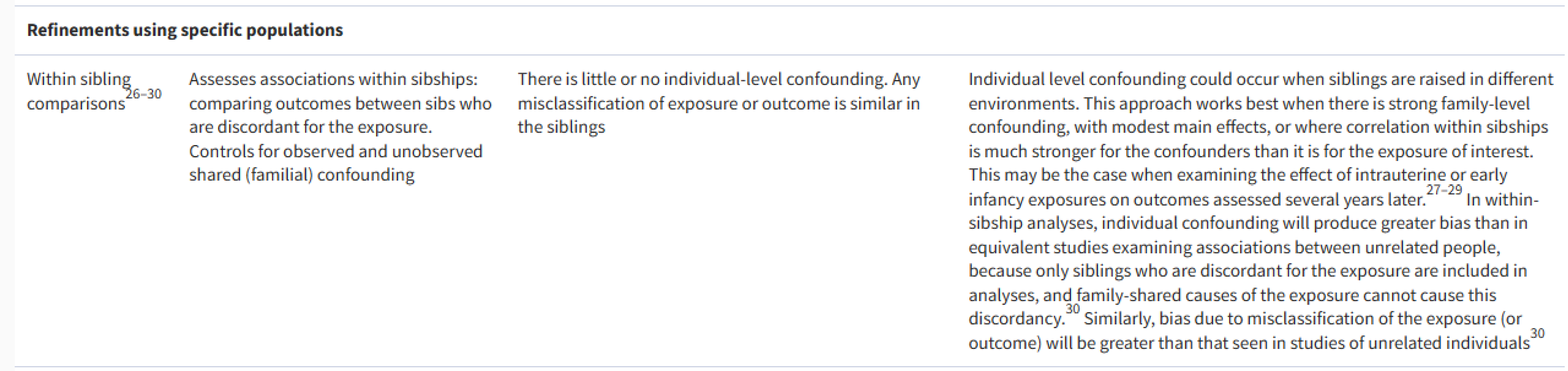


Table 1.

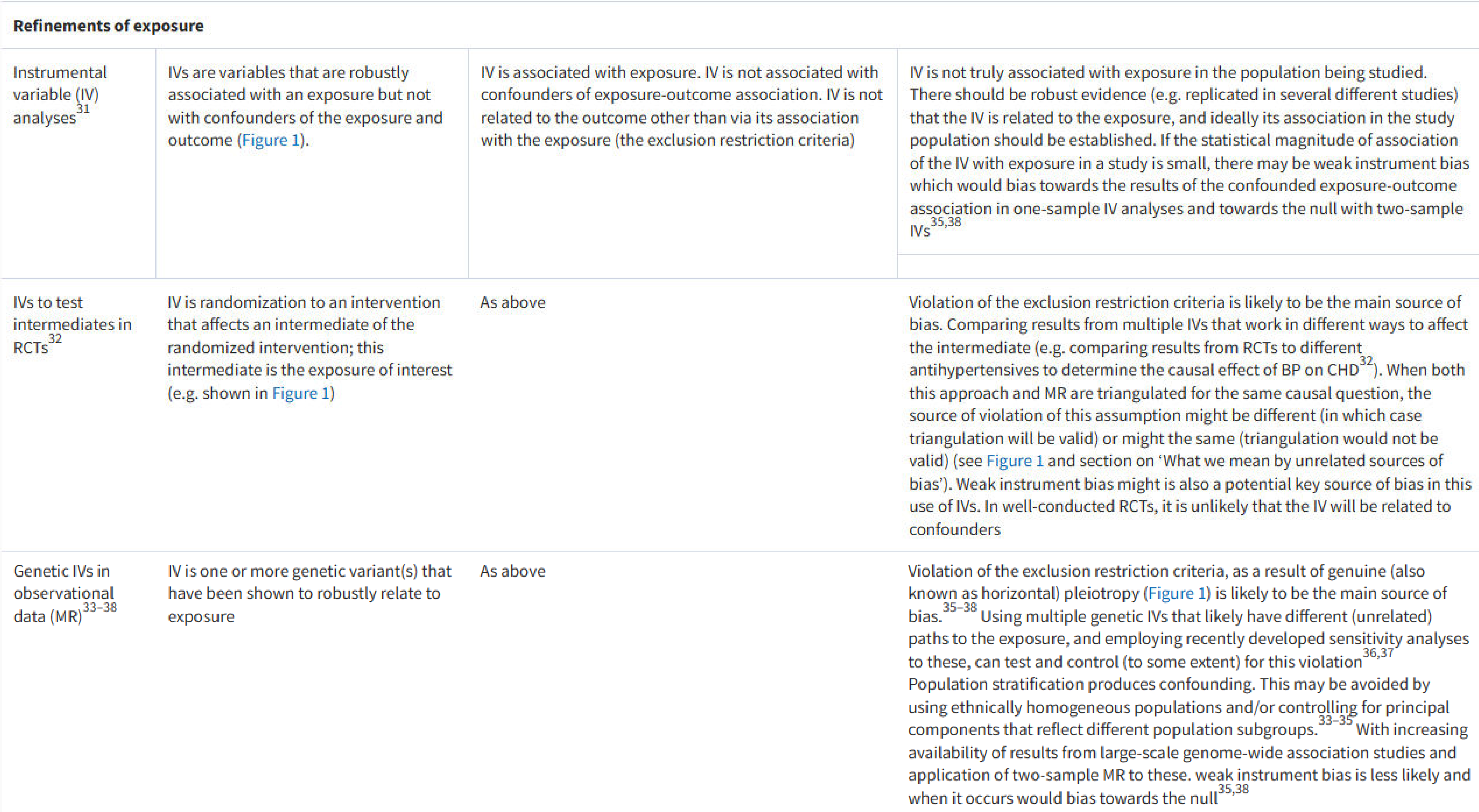
Key sources of bias in different aetiological epidemiology approaches

Approach	Description	Assumptions	General key sources of potential bias[a]
Conventional approaches			
Randomized controlled trials[17,18]	Prospective intervention study in which people are randomly allocated to comparison groups that are given different interventions	Intervention groups are similar with the exception of the intervention	Lack of concealment of the random allocation. Failure to maintain the original randomized status of participants when comparing outcomes and lack of blinding to which group participants have been randomised. Differential loss to follow-up, for example due to adverse effects of the intervention or a perception that there is no benefit
Multivariable regression in observational data[19,20]	The application of multivariable regression to observational data	No residual confounding (all confounders are accurately measured and controlled for). Participants are not selected to participate or to be included in analyses in a way that produces a spurious associations. Any misclassification of exposure is not related to the outcome, and vice versa, and misclassification of covariables are not systematically related to outcome or exposure	Unmeasured or poorly measured confounders (residual confounding). Reverse causality. Misclassification of exposure is related to the outcome (or vice versa). Differential missing data between exposure levels, for example due loss to follow-up in prospective cohort studies or reporting bias in case-control studies

Cross-context comparisons[21]	Compares results between two or more populations in different contexts that result in confounding structures being different	Different results between populations are due to different confounding structures and not due to true differences in causal effects between populations. Similar results between populations cannot be explained by confounding, given the differences between the populations/contexts in their confounding structures. There are no other (than confounding) sources of bias that could explain similar or different results between the two populations	Confounders are, in fact, the same in the populations being compared. For observed confounders, differences between the two populations should be established. There are different sources of bias (over and above different confounding structures), for example differential misclassification of exposure or outcome that investigators are unaware of. Measurement of the exposure and outcome and the quality of these measurements should be the same or very similar in the populations being compared
Different control groups[22–24]	Use of two or more different control groups in a case-control study, where the bias for the different control groups is expected to be in different directions	The different sources of bias for the different control groups are different and would produce different results	If biases are in fact the same in the different control groups being compared, the inference made when comparing them will be misleading. If there are different sources of bias between the different control groups, but these nonetheless distort the finding in the same direction, this will also be misleading. Inference may be incorrect, if there are a priori incorrect assumptions about one of the control groups being least biased for the specific research question. This may be less statistically efficient than having just one control group, as with fixed resources it would imply using a smaller number of controls for each of the two groups (and possibly a smaller number of cases, as resources would be required to recruit two different sources of controls)
Natural experiments[25]	Populations are compared in the belief that biases, such as confounding structures are similar between them. One (or more) of the populations has had a 'natural' exposure or are 'quasi-randomly' exposed. Natural exposure, e.g. flood or famine; quasi-randomization, e.g. those resulting from different timing of introduction to policies, such as smoking bans in public places	The populations being compared are similar with the exception of the naturally or quasi-randomized exposure	Populations differ on characteristics that confound the association. Misclassification of the outcome is related to the naturally occurring exposure. Ideally, identical methods for measuring the outcome should be used in each population. If associations are measured at the aggregated population level but interpreted as if they apply to individuals within the population, there may be bias due to the ecological fallacy



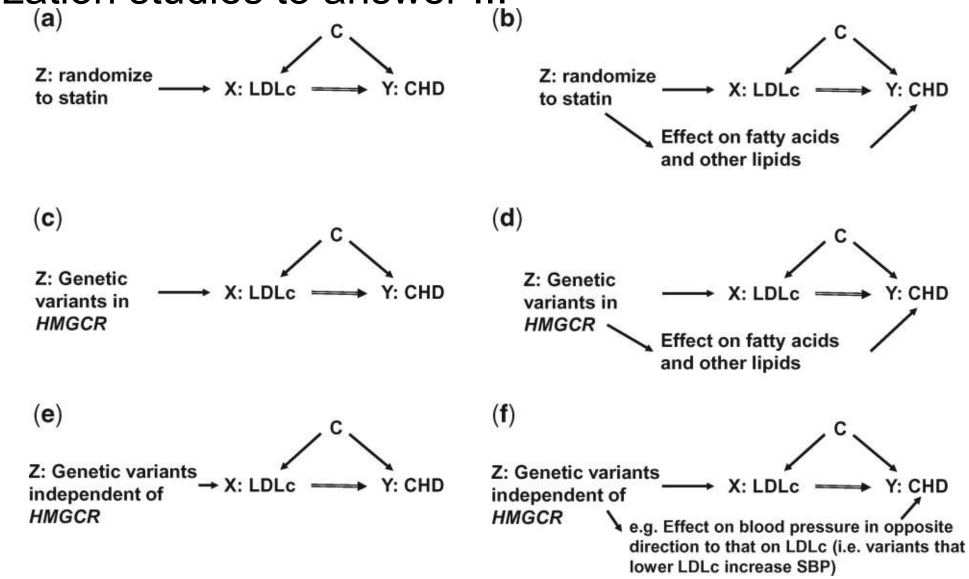
Refinements using specific populations			
Within sibling comparisons[26–30]	Assesses associations within sibships: comparing outcomes between sibs who are discordant for the exposure. Controls for observed and unobserved shared (familial) confounding	There is little or no individual-level confounding. Any misclassification of exposure or outcome is similar in the siblings	Individual level confounding could occur when siblings are raised in different environments. This approach works best when there is strong family-level confounding, with modest main effects, or where correlation within sibships is much stronger for the confounders than it is for the exposure of interest. This may be the case when examining the effect of intrauterine or early infancy exposures on outcomes assessed several years later.[27–29] In within-sibship analyses, individual confounding will produce greater bias than in equivalent studies examining associations between unrelated people, because only siblings who are discordant for the exposure are included in analyses, and family-shared causes of the exposure cannot cause this discordancy.[30] Similarly, bias due to misclassification of the exposure (or outcome) will be greater than that seen in studies of unrelated individuals[30]



Refinements of exposure			
Instrumental variable (IV) analyses[31]	IVs are variables that are robustly associated with an exposure but not with confounders of the exposure and outcome (Figure 1).	IV is associated with exposure. IV is not associated with confounders of exposure-outcome association. IV is not related to the outcome other than via its association with the exposure (the exclusion restriction criteria)	IV is not truly associated with exposure in the population being studied. There should be robust evidence (e.g. replicated in several different studies) that the IV is related to the exposure, and ideally its association in the study population should be established. If the statistical magnitude of association of the IV with exposure in a study is small, there may be weak instrument bias which would bias towards the results of the confounded exposure-outcome association in one-sample IV analyses and towards the null with two-sample IVs[35,38]
IVs to test intermediates in RCTs[32]	IV is randomization to an intervention that affects an intermediate of the randomized intervention; this intermediate is the exposure of interest (e.g. shown in Figure 1)	As above	Violation of the exclusion restriction criteria is likely to be the main source of bias. Comparing results from multiple IVs that work in different ways to affect the intermediate (e.g. comparing results from RCTs to different antihypertensives to determine the causal effect of BP on CHD[32]). When both this approach and MR are triangulated for the same causal question, the source of violation of this assumption might be different (in which case triangulation will be valid) or might the same (triangulation would not be valid) (see Figure 1 and section on 'What we mean by unrelated sources of bias'). Weak instrument bias might is also a potential key source of bias in this use of IVs. In well-conducted RCTs, it is unlikely that the IV will be related to confounders
Genetic IVs in observational data (MR)[33–38]	IV is one or more genetic variant(s) that have been shown to robustly relate to exposure	As above	Violation of the exclusion restriction criteria, as a result of genuine (also known as horizontal) pleiotropy (Figure 1) is likely to be the main source of bias.[35–38] Using multiple genetic IVs that likely have different (unrelated) paths to the exposure, and employing recently developed sensitivity analyses to these, can test and control (to some extent) for this violation[36,37] Population stratification produces confounding. This may be avoided by using ethnically homogeneous populations and/or controlling for principal components that reflect different population subgroups.[33–35] With increasing availability of results from large-scale genome-wide association studies and application of two-sample MR to these. weak instrument bias is less likely and when it occurs would bias towards the null[35,38]

Non-genetic IVs in observational data[25]	IV is non-genetic, examples include use of exposures in other family members as IVs for the index participants' exposure, or a 'natural' occurring phenomenon (such as famine or flood); this approach is commonly used in natural experiments[25]	As above	Association of the IV with confounders of the exposure-outcome association are more likely with this approach than IVs for intermediates in an RCT or MR. Violation of the exclusion restriction criteria is possible; given the wide range of non-genetic IVs that could potentially be used, the extent to which this may be a major source of bias is hard to state in a general way. Weak instrument bias is possible.
Exposure negative control studies[39–42]	Aims to reproduce the same conditions as the 'real' study, but uses a different (negative control) exposure that is not plausibly causally related to outcome	The key sources of bias, including specific confounders, misclassification bias and other biases, are the same for the real and negative control exposures. The negative control exposure does not have a causal effect on the real outcome. To sensibly compare the real and negative control exposure, they should ideally be similarly scaled. This should be possible when negative control exposures are used to test critical or sensitive periods (see section on duration and timing of exposure being assessed with different exposures)	There are differences in the sources of bias between the real and negative control exposure. Attempts to explore this (e.g. exploring the association of observed confounders with the negative control exposure) should be made. There is a real (but unknown) causal effect of the negative control exposure on the outcome
Refinements of outcome			
Outcome negative control studies[39–42]	As above, except here a different outcome is selected for the negative control study	As above, except here a different outcome is selected for the negative control study	As above, if either assumption is violated there could be biased inference from the comparison of the real with the negative control study

Figure 1. Illustrative example of instrumental variable analyses in RCTs and Mendelian randomization studies to answer ...


(a)
C
Z: randomize to statin
X: LDLc
Y: CHD
(b)
C
Z: randomize to statin
X: LDLc
Y: CHD
Effect on fatty acids and other lipids
(c)
C
Z: Genetic variants in *HMGCR*
X: LDLc
Y: CHD
(d)
C
Z: Genetic variants in *HMGCR*
X: LDLc
Y: CHD
Effect on fatty acids and other lipids
(e)
C
Z: Genetic variants independent of *HMGCR*
X: LDLc
Y: CHD
(f)
C
Z: Genetic variants independent of *HMGCR*
X: LDLc
Y: CHD
e.g. Effect on blood pressure in opposite direction to that on LDLc (i.e. variants that lower LDLc increase SBP)

Int J Epidemiol, Volume 45, Issue 6, December 2016, Pages 1866–1886, https://doi.org/10.1093/ije/dyw314

- Definition: consilience, i.e. coherence of different approaches
- Framing: scientific reproducibility
- Relationship with causal inference: design ≡ directed acyclic graph
- How it works

How it works ?

By identification of different sources of evidence

By sifting likely sources and direction of bias operating within each

By assessing duration and timing of exposure of different approaches

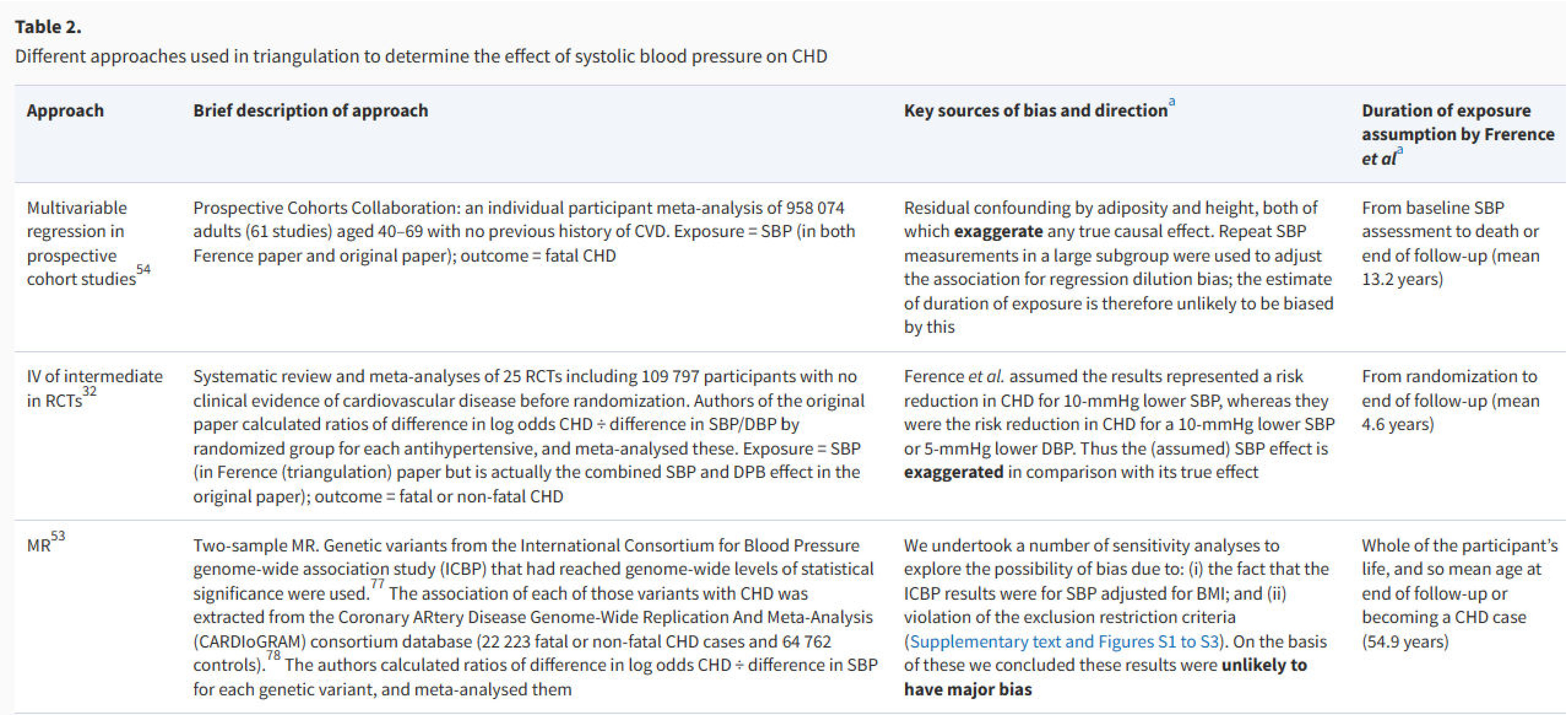
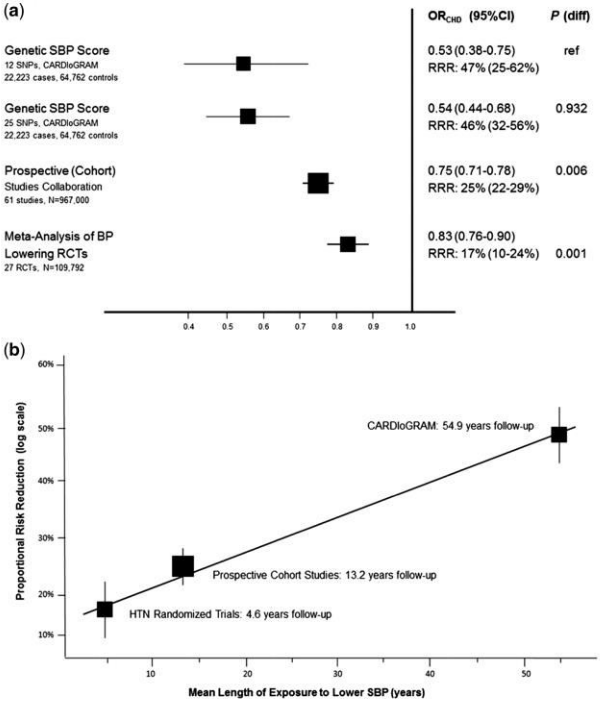


Table 2.

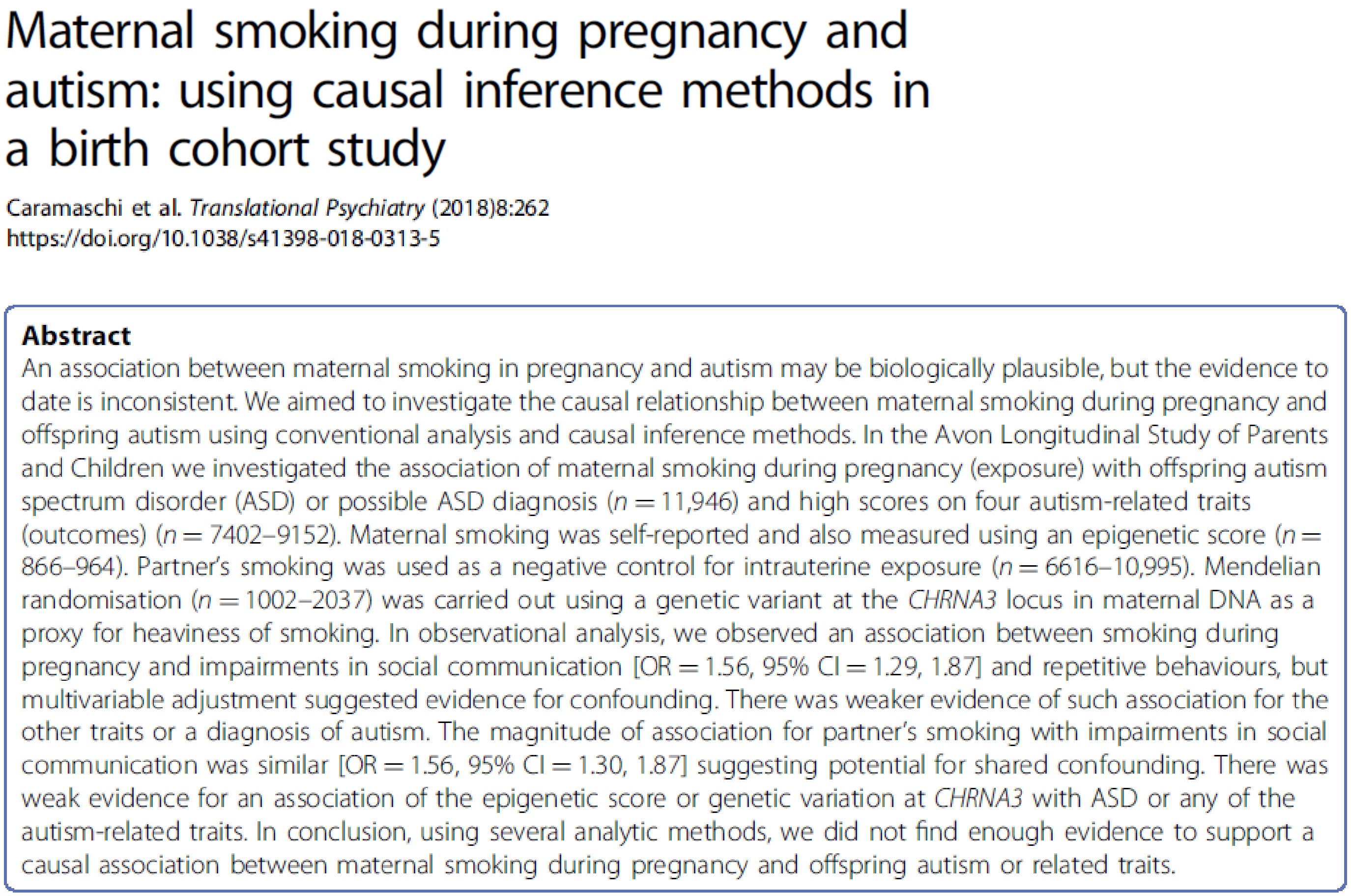
Different approaches used in triangulation to determine the effect of systolic blood pressure on CHD

Approach	Brief description of approach	Key sources of bias and direction[a]	Duration of exposure assumption by Frerence *et al*[a]
Multivariable regression in prospective cohort studies[54]	Prospective Cohorts Collaboration: an individual participant meta-analysis of 958 074 adults (61 studies) aged 40–69 with no previous history of CVD. Exposure = SBP (in both Ference paper and original paper); outcome = fatal CHD	Residual confounding by adiposity and height, both of which **exaggerate** any true causal effect. Repeat SBP measurements in a large subgroup were used to adjust the association for regression dilution bias; the estimate of duration of exposure is therefore unlikely to be biased by this	From baseline SBP assessment to death or end of follow-up (mean 13.2 years)
IV of intermediate in RCTs[32]	Systematic review and meta-analyses of 25 RCTs including 109 797 participants with no clinical evidence of cardiovascular disease before randomization. Authors of the original paper calculated ratios of difference in log odds CHD ÷ difference in SBP/DBP by randomized group for each antihypertensive, and meta-analysed these. Exposure = SBP (in Ference (triangulation) paper but is actually the combined SBP and DPB effect in the original paper); outcome = fatal or non-fatal CHD	Ference *et al.* assumed the results represented a risk reduction in CHD for 10-mmHg lower SBP, whereas they were the risk reduction in CHD for a 10-mmHg lower SBP or 5-mmHg lower DBP. Thus the (assumed) SBP effect is **exaggerated** in comparison with its true effect	From randomization to end of follow-up (mean 4.6 years)
MR[53]	Two-sample MR. Genetic variants from the International Consortium for Blood Pressure genome-wide association study (ICBP) that had reached genome-wide levels of statistical significance were used.[77] The association of each of those variants with CHD was extracted from the Coronary ARtery Disease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) consortium database (22 223 fatal or non-fatal CHD cases and 64 762 controls).[78] The authors calculated ratios of difference in log odds CHD ÷ difference in SBP for each genetic variant, and meta-analysed them	We undertook a number of sensitivity analyses to explore the possibility of bias due to: (i) the fact that the ICBP results were for SBP adjusted for BMI; and (ii) violation of the exclusion restriction criteria (Supplementary text and Figures S1 to S3). On the basis of these we concluded these results were **unlikely to have major bias**	Whole of the participant's life, and so mean age at end of follow-up or becoming a CHD case (54.9 years)

Figure 2. Triangulation of effect of systolic blood pressure on CHD risk from three approaches (RCT, multivariable ...



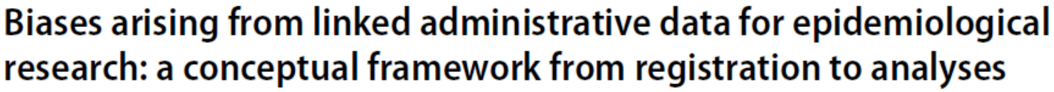
Int J Epidemiol, Volume 45, Issue 6, December 2016, Pages 1866–1886, https://doi.org/10.1093/ije/dyw314



Maternal smoking during pregnancy and autism: using causal inference methods in a birth cohort study

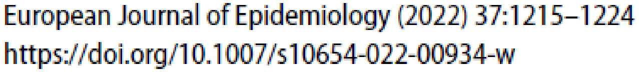
Caramaschi et al. *Translational Psychiatry* (2018)8:262
https://doi.org/10.1038/s41398-018-0313-5

Abstract
An association between maternal smoking in pregnancy and autism may be biologically plausible, but the evidence to date is inconsistent. We aimed to investigate the causal relationship between maternal smoking during pregnancy and offspring autism using conventional analysis and causal inference methods. In the Avon Longitudinal Study of Parents and Children we investigated the association of maternal smoking during pregnancy (exposure) with offspring autism spectrum disorder (ASD) or possible ASD diagnosis ($n = 11,946$) and high scores on four autism-related traits (outcomes) ($n = 7402$–9152). Maternal smoking was self-reported and also measured using an epigenetic score ($n = 866$–964). Partner's smoking was used as a negative control for intrauterine exposure ($n = 6616$–$10,995$). Mendelian randomisation ($n = 1002$–2037) was carried out using a genetic variant at the *CHRNA3* locus in maternal DNA as a proxy for heaviness of smoking. In observational analysis, we observed an association between smoking during pregnancy and impairments in social communication [OR = 1.56, 95% CI = 1.29, 1.87] and repetitive behaviours, but multivariable adjustment suggested evidence for confounding. There was weaker evidence of such association for the other traits or a diagnosis of autism. The magnitude of association for partner's smoking with impairments in social communication was similar [OR = 1.56, 95% CI = 1.30, 1.87] suggesting potential for shared confounding. There was weak evidence for an association of the epigenetic score or genetic variation at *CHRNA3* with ASD or any of the autism-related traits. In conclusion, using several analytic methods, we did not find enough evidence to support a causal association between maternal smoking during pregnancy and offspring autism or related traits.

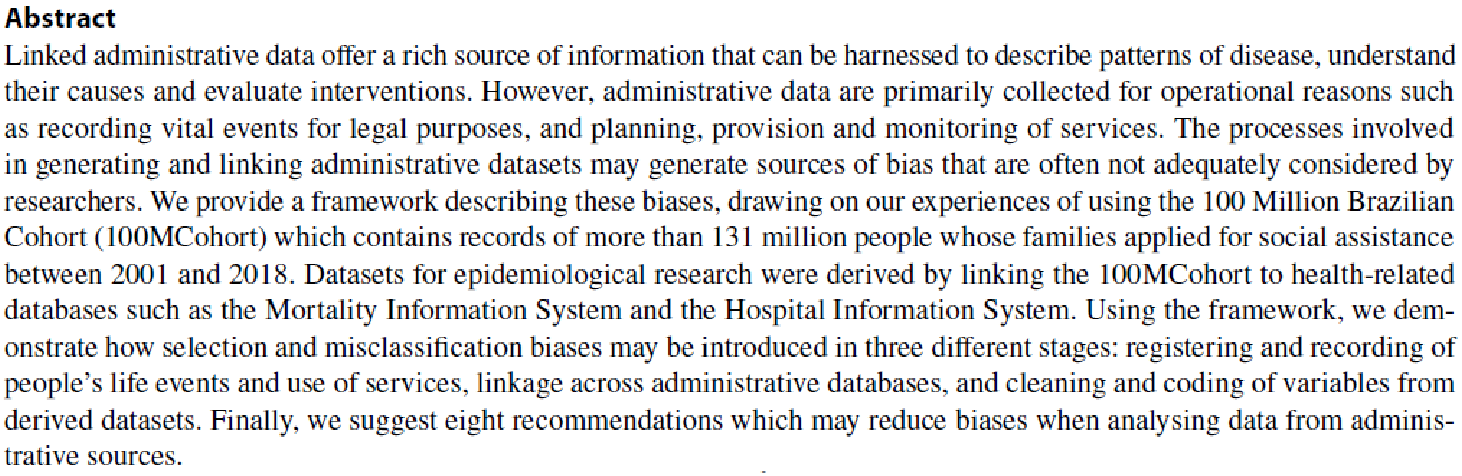


Biases arising from linked administrative data for epidemiological research: a conceptual framework from registration to analyses

Shaw *et al.*

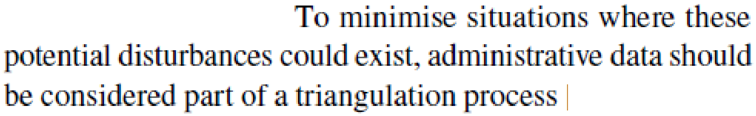


European Journal of Epidemiology (2022) 37:1215–1224
https://doi.org/10.1007/s10654-022-00934-w



Abstract

Linked administrative data offer a rich source of information that can be harnessed to describe patterns of disease, understand their causes and evaluate interventions. However, administrative data are primarily collected for operational reasons such as recording vital events for legal purposes, and planning, provision and monitoring of services. The processes involved in generating and linking administrative datasets may generate sources of bias that are often not adequately considered by researchers. We provide a framework describing these biases, drawing on our experiences of using the 100 Million Brazilian Cohort (100MCohort) which contains records of more than 131 million people whose families applied for social assistance between 2001 and 2018. Datasets for epidemiological research were derived by linking the 100MCohort to health-related databases such as the Mortality Information System and the Hospital Information System. Using the framework, we demonstrate how selection and misclassification biases may be introduced in three different stages: registering and recording of people's life events and use of services, linkage across administrative databases, and cleaning and coding of variables from derived datasets. Finally, we suggest eight recommendations which may reduce biases when analysing data from administrative sources.



To minimise situations where these potential disturbances could exist, administrative data should be considered part of a triangulation process

Conclusions

- Reproducibility may be a “no issue”

  we see a reproducibility problem because we look at with a wrong glass

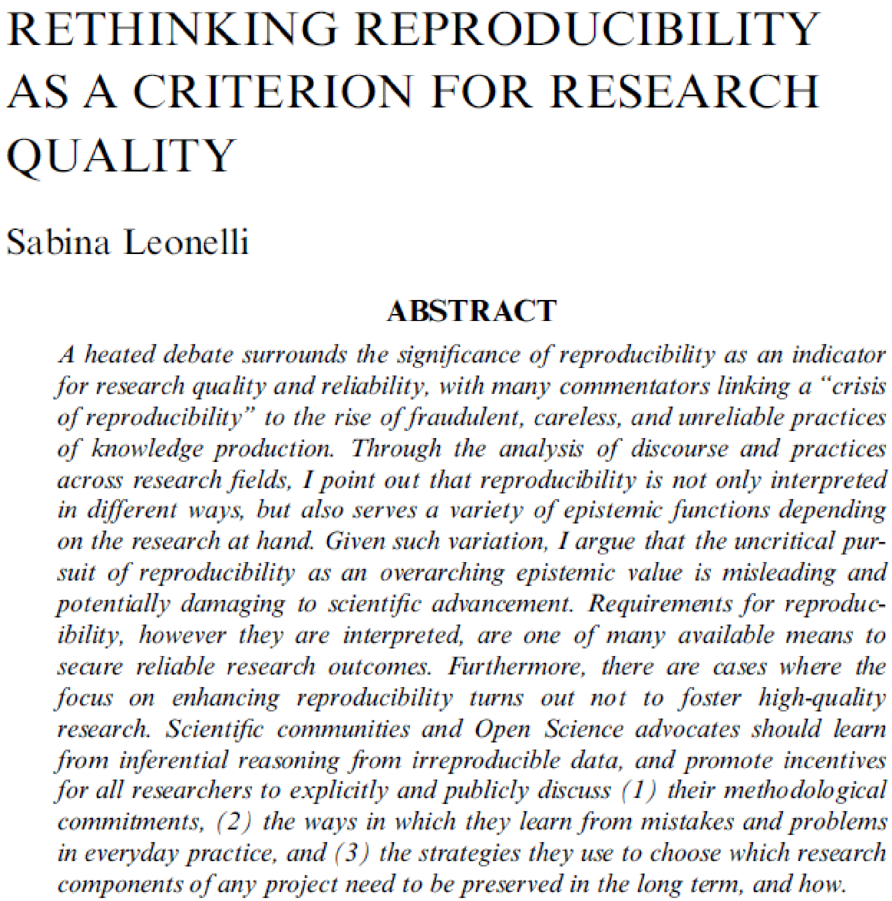
- Uncertainty, in its complexity, is not solvable

- Scientific reproducibility is controversial

  we are urged to take action and as reported in the history of the WHO International Agency for Research on Cancer (Saracci Wild 2015)

  “… the aim was to develop an instrument capable of evaluating the best evidence available at a given time …”

  of course this aim cannot be reached internally



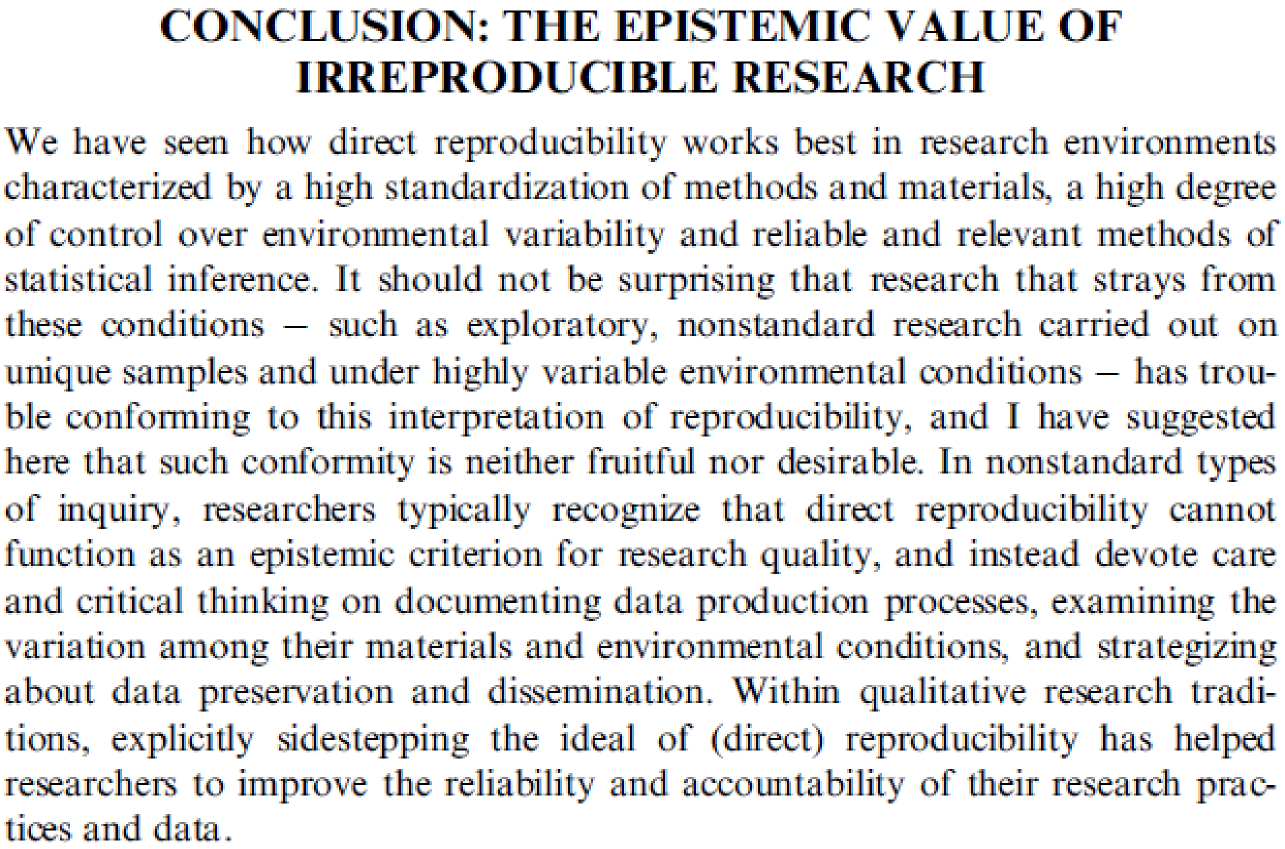
RETHINKING REPRODUCIBILITY AS A CRITERION FOR RESEARCH QUALITY

Sabina Leonelli

ABSTRACT

A heated debate surrounds the significance of reproducibility as an indicator for research quality and reliability, with many commentators linking a "crisis of reproducibility" to the rise of fraudulent, careless, and unreliable practices of knowledge production. Through the analysis of discourse and practices across research fields, I point out that reproducibility is not only interpreted in different ways, but also serves a variety of epistemic functions depending on the research at hand. Given such variation, I argue that the uncritical pursuit of reproducibility as an overarching epistemic value is misleading and potentially damaging to scientific advancement. Requirements for reproducibility, however they are interpreted, are one of many available means to secure reliable research outcomes. Furthermore, there are cases where the focus on enhancing reproducibility turns out not to foster high-quality research. Scientific communities and Open Science advocates should learn from inferential reasoning from irreproducible data, and promote incentives for all researchers to explicitly and publicly discuss (1) their methodological commitments, (2) the ways in which they learn from mistakes and problems in everyday practice, and (3) the strategies they use to choose which research components of any project need to be preserved in the long term, and how.

Sabina Leonelli, "Rethinking Reproducibility as a Criterion for Research Quality"
In Including a Symposium on Mary Morgan:
Curiosity, Imagination, and Surprise. 05 Oct 2018; 129-146.
https://doi.org/10.1108/S0743-41542018000036B009

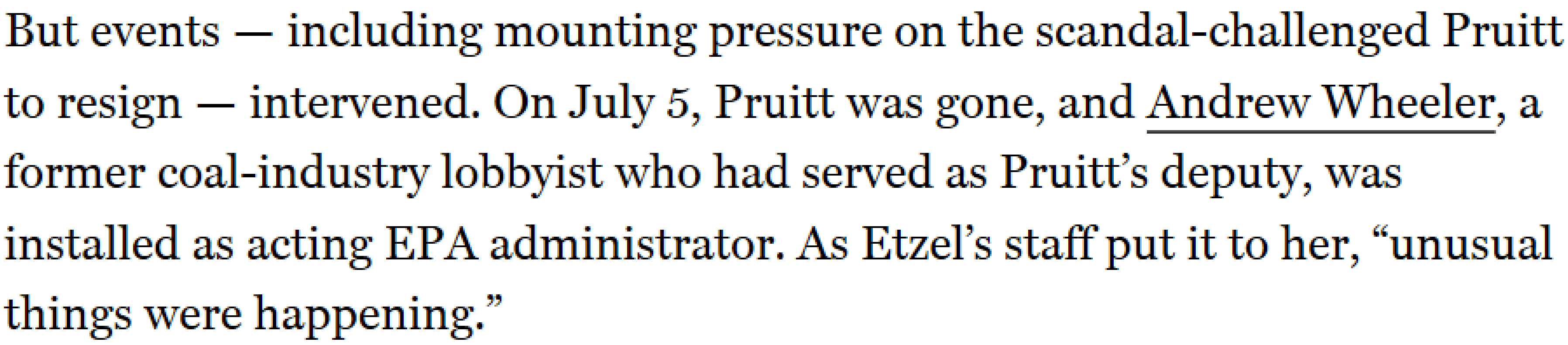


CONCLUSION: THE EPISTEMIC VALUE OF IRREPRODUCIBLE RESEARCH

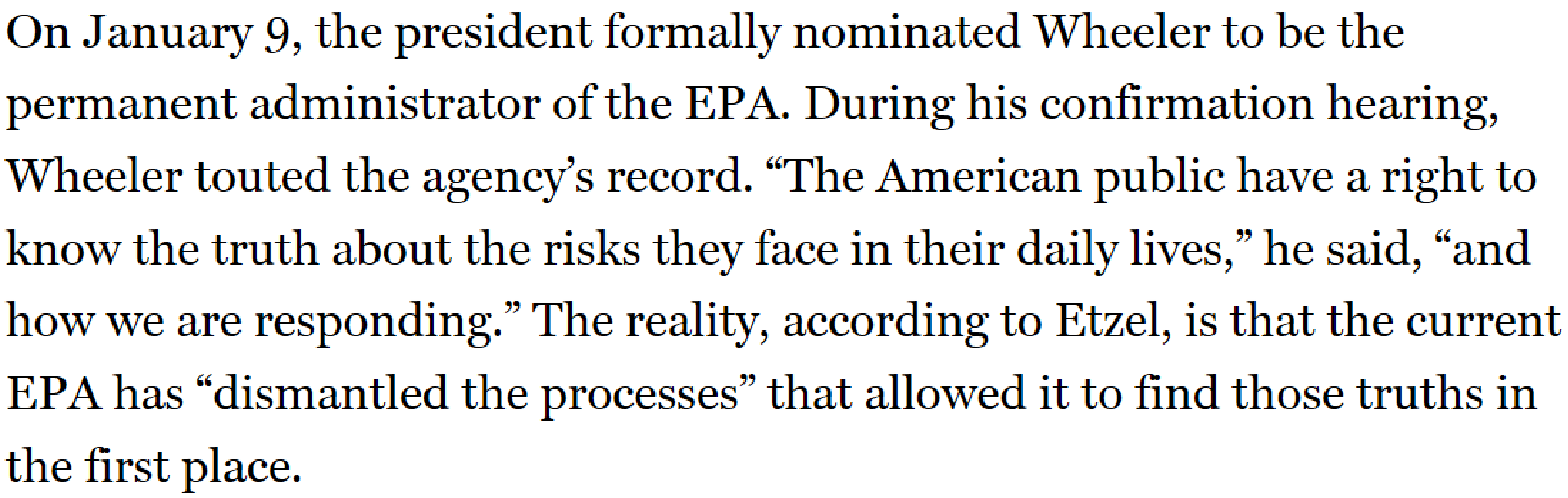
We have seen how direct reproducibility works best in research environments characterized by a high standardization of methods and materials, a high degree of control over environmental variability and reliable and relevant methods of statistical inference. It should not be surprising that research that strays from these conditions – such as exploratory, nonstandard research carried out on unique samples and under highly variable environmental conditions – has trouble conforming to this interpretation of reproducibility, and I have suggested here that such conformity is neither fruitful nor desirable. In nonstandard types of inquiry, researchers typically recognize that direct reproducibility cannot function as an epistemic criterion for research quality, and instead devote care and critical thinking on documenting data production processes, examining the variation among their materials and environmental conditions, and strategizing about data preservation and dissemination. Within qualitative research traditions, explicitly sidestepping the ideal of (direct) reproducibility has helped researchers to improve the reliability and accountability of their research practices and data.

(Leonelli, cit)

http://nymag.com/intelligencer/2019/02/trump-epa-risking-health-of-american-children.html



But events — including mounting pressure on the scandal-challenged Pruitt to resign — intervened. On July 5, Pruitt was gone, and Andrew Wheeler, a former coal-industry lobbyist who had served as Pruitt's deputy, was installed as acting EPA administrator. As Etzel's staff put it to her, "unusual things were happening."



On January 9, the president formally nominated Wheeler to be the permanent administrator of the EPA. During his confirmation hearing, Wheeler touted the agency's record. "The American public have a right to know the truth about the risks they face in their daily lives," he said, "and how we are responding." The reality, according to Etzel, is that the current EPA has "dismantled the processes" that allowed it to find those truths in the first place.

Replicability Issues in Epidemiologic Research

Annibale Biggeri

Department of Cardiac, Thoracic, Vascular Sciences and Public Health

Unit of Biostatistics, Epidemiology and Public Health

University of Padua

annibale.biggeri@unipd.it