Sterne, J. (2018). Commentary: does the selective inversion approach demonstrate bias in the results of studies using routinely collected data? *BMJ*, *362*, [k3259]. https://doi.org/10.1136/bmj.k3259

To coin a phrase: did the selective inversion approach demonstrate bias in the results of studies using routinely collected data?

Jonathan Sterne, Professor of Medical Statistics and Epidemiology, NIHR Bristol Biomedical Research Centre, Department of Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, Bristol, BS8 2BN, UK.

Acknowledgement: This commentary draws on comments made by the late Professor Doug Altman, in his capacity as *BMJ* statistical editor.

Standfirst: Selective inversion ("coining") of the direction of comparisons led to strong disagreements over the validity of a paper examining bias in studies of interventions based on routinely collected data, and a correction published in this week's *BMJ*.

In early 2016 *BMJ* published a paper by Hemkens et al. entitled "Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey".[1] This study used a novel design: the authors identified observational studies based on routinely collected data (RCD) that were performed before any randomized controlled trial (RCT) on the same clinical question, reported on all-cause mortality, and used a propensity score-based analysis to control confounding. The authors identified 16 RCD studies for which results from RCTs were subsequently reported. The main result was: "Overall, RCD studies showed significantly more favorable mortality estimates by 31% than subsequent trials (summary relative odds ratio [ROR] 1.31 (95% CI 1.03 to 1.65; I^2=0%))." The authors also reported that "Confidence intervals in nine (56%) RCD studies did not include the RCT effect estimate." They concluded that "studies that use routinely collected health data frequently do not agree with subsequent randomized trials" and that "the emerging pattern was that RCD studies systematically and substantially overestimated the mortality benefits of medical treatments compared with subsequent trials investigating the same question".

In August 2016 Franklin et al. submitted to *BMJ* a paper that criticised the methods used by Hemkens et al. After *BMJ* invited these authors to submit their comments as a rapid response rather than a separate paper, the work was published in *Epidemiological Methods*.[2] Hemkens et al. had "inverted the RCD effect estimates where necessary so that each RCD study indicated an odds ratio less than 1"[1]. Their aim was to overcome the problem that some of the pairs compared two active interventions, so that the choice of direction of comparison was arbitrary. Franklin et al. showed that, under reasonable assumptions, the selective inversion approach employed by Hemkens et al. will lead the average ROR across studies to be greater than 1, even if there is no systematic difference between results from RCD studies and those from subsequent RCTs. This is because the sampling error in the OR for the RCD study is related both to the ROR and whether the OR is greater than 1 and so is inverted.

Publication of the paper by Franklin et al. was noted in a rapid response posted on bmj.com by Dr Merino (*BMJ* US Research Editor) in August 2017. In September 2017 Hemkens et al. posted a further rapid response that dismissed the arguments and analyses of Franklin et al. as "flawed and misleading". In October 2017 Franklin et al. responded that Hemkens et al. had not addressed the core statistical problem with their analyses. In November 2017 Dr Merino acknowledged receipt of a letter calling for retraction of the paper by Hemkens et al. Further comments by these authors and others were posted in a series of rapid responses on bmj.com.

Hemkens et al. argued that their selective inversion approach, which they described as "coining", was reasonable because they had taken the perspective of clinicians facing new evidence, having no randomized trials, and having to decide whether they use a new promising treatment. A counter-argument is that, regardless of the clinical question being addressed, an appropriate method should result in an average ROR of 1 (corresponding to results in the two types of study being the same), in the absence of systematic differences between RCDs and RCTs. The "coining" approach has been

used in a number of publications, dating back at least to 2005[3]. It seems possible that the bias that was demonstrated by Franklin et al. affected some of these publications.

Extensive discussions, both on bmj.com and between Hemkens et al. and BMJ editors and reviewers, have now led to publication of a correction to the paper by Hemkens et al.[4] These authors now present revised analyses that avoid selective inversion, either by restriction to comparisons of an active treatment with a control that did not receive this treatment, or by classifying treatments as "control" if they represented the established treatment or that shown in RCTs to work in the relevant indication. Detailed justifications for these classifications are provided in the correction. In general, the conclusions from the re-analyses are not greatly changed compared with the original paper. Some of the estimated RORs move further away from the null value of 1, while some of the confidence intervals become wider. In a combined analysis of the 15 clinical questions for which a direction could be chosen, the ROR was 1.38 (95% CI 1.09-1.74, I^2=0%). The similarity of these results to those of the original study is because the direction of the results was recoded for only three.

Results of RCD studies can be affected by both confounding and by other biases that are avoided in high quality RCTs. Randomization, providing that it is properly done, protects against confounding bias (other than through chance differences between intervention and control that are accounted for in the 95% confidence interval for the result)[5]. Even the most ardent advocates of RCD studies would agree that it is not always possible to control confounding (whether using propensity scores, as in the RCD studies considered by Hemkens et al., or any other method). The rapid response by Suisa et al. on 1 April 2016 identified additional biases (such as immortal time bias) that may have affected results of particular RCDs included in the study of Hemkens et al. It seems implausible that the true value of the ROR for each of the comparisons made by Hemkens et al. is 1, given that results of each RCD and each RCT will be at differing risks of bias. However, the magnitude and direction of of bias is likely to vary between studies and settings, leading to increases in between-study heterogeneity.[6]

Hemkens et al. are to be congratulated on their novel study design, and their attempt to quantify differences between results of RCD studies and subsequently published RCTs. It is desirable, as they note in their correction, that their findings are replicated in a new study. Such a study should have a pre-specified analysis plan, with classification of "experimental" and "control" interventions based on clear rules and classified blind to the direction and magnitude of the difference between the results of the RCD and RCT. The revised analyses presented in the correction by Hemkens et al. were inevitably post hoc, and the need for clear criteria is illustrated by the study of on-pump (with cardiopulmonary bypass [CPB]) versus off-pump (without CPB) cardiac surgery by Karthik et al.[7] Hemkens et al. have changed the direction of comparison to consider off-pump surgery as control. However, it would seem more appropriate to have considered off-pump cardiac surgery as the new approach: Karthik et al. stated "In the last 15 years, CABG without CPB has developed into a viable alternative to on-pump CABG".

I have some areas of disagreement with the correction by Hemkens et al. Most importantly, they continue to describe the difference between the analyses that are and are not based on selective inversion as related to the clinical question of interest. Franklin et al demonstrated the selective inversion approach to be biased: it should not be used in future analyses of empirical studies. Second, the similarity of the results of the re-analysis with those in the original paper does not seem "remarkable", given that the direction of comparison was changed for few RCD-RCT pairs.

Third, Hemkens et al. reported on the proportion of RCD-RCT pairs for which the confidence interval for the RCD estimate did not include the estimated effect from RCTs: "in more than half of cases (nine of 16; 56%), the confidence intervals of the RCD based estimate did not include the mortality effect found in subsequent randomized trial evidence", and concluded that "studies that use routinely collected health data frequently do not agree with subsequent randomized trials". However, the precision of estimates based on RCTs was often much less than the precision of

estimates based on RCD studies. There is only a small probability that an imprecisely estimated RCT effect lies within the 95% CI for a precisely estimated RCD effect: see for example the comparison of hemodialysis with peritoneal dialysis in renal replacement therapy by Winkelmayer et al.[8] It is not possible to interpret such 'disagreement' in the absence of a clear statement of the expected proportion of topics for which it is expected under the null hypothesis. Franklin et al. suggested an alternative based on comparing observed and expected overlap between confidence intervals, which Hemkens et al. described as "notoriously underpowered" and "clinically irrelevant." In their correction, they discuss issues of power and Type I and II error rates associated with comparing expected versus observed overlap of CIs. I regret that they did not clarify levels of expected discrepancy associated with their own approach.

This has been a bruising process for many of those concerned. Nonetheless, it has resulted in important clarifications about methods for empirical research on biases associated with study designs and characteristics. Such research has underpinned development of both reporting guidelines[9] and tools to assess risk of bias[10], and is of particular importance given the opportunities arising from availability of very large and detailed routine health care datasets, and the risks that analyses of such data lead to precise but misleading estimates of the effect of interventions.

Jonathan Sterne, 25 July 2018

1. Hemkens LG, Contopoulos-Ioannidis DG and Ioannidis JPA. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. BMJ 2016;352:i493. http://dx.doi.org/10.1136/bmj.i493
2. Franklin JM, Dejene S, Huybrechts KF, Wang SV, Kulldorff M and Rothman KJ. A Bias in the Evaluation of Bias Comparing Randomized Trials with Nonexperimental Studies. Epidemiological Methods 2017; 20160018. DOI: 10.1515/em-2016-0018
3. Ioannidis JPA and Trikalinos TA. Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. Journal of Clinical Epidemiology 2005; 58(6): 543–549. https://doi.org/10.1016/j.jclinepi.2004.10.019.
4. Hemkens LG, Contopoulos-Ioannidis DG and Ioannidis JPA. Re-analysis: Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials. BMJ 2018, in press.
5. Senn S. Seven myths of randomisation in clinical trials. Statistics in Medicine 2013; 32: 1439–1450. DOI: 10.1002/sim.5713
6. Savović J, Jones HE, Altman DG, Harris RJ, Jűni P, Pildal J, Als-Nielsen B, Balk EM, Gluud C, Gluud LL, Ioannidis JPA, Schulz KF, Beynon R, Welton NJ, Wood L, Moher D, Deeks JJ and Sterne JAC. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. Health Technology Assessment 2012; 16: 35.
7. Karthik S, Musleh G, Grayson AD, et al. Effect of avoiding cardiopulmonary bypass in non-elective coronary artery bypass surgery: a propensity score analysis. Eur J Cardiothorac Surg 2003; 24: 66-71.
8. Winkelmayer WC, Glynn RJ, Mittleman MA, Levin R, Pliskin JS, Avorn J. Comparing mortality of elderly patients on hemodialysis versus peritoneal dialysis: a propensity score approach. Journal of the American Society of Nephrology 2002; 13: 2353-62.
9. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M and Altman DG. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c869. DOI: https://doi.org/10.1136/bmj.c869
10. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savović J, Schulz KF, Weeks L, Sterne JAC, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomized trials. BMJ 2011; 343: d5928. DOI: 10.1136/bmj.d5928.