5. Croke K, Hicks JH, Hsu E, Kremer M, Miguel E. *Does Mass Deworming Affect Child Nutrition? Meta-analysis, Cost-effectiveness, and Statistical Power*. 2016. http://scholar.harvard.edu/files/kcroke/files/ug_lr_deworming_071714.pdf (2 November 2016,date last accessed).
6. Miguel E, Camerer C, Casey K, *et al*. Promoting transparency in social science research. *Science* 2014;**343:**30–31.
7. Joseph SA, Casapia M, Montresor A *et al*. The effect of deworming on growth in one-year-old children living in a soil-transmitted helminth-endemic area of Peru: A randomized controlled trial. *PLoS Negl Trop Dis* 2015;**9:**e0004020.
8. Joseph SA, Casapia M, Rahme E, Pezo L, Blouin B, Gyorkos TW. The effect of deworming on early childhood development in Peru: A randomized controlled trial. *PloS Negl Trop Dis* 2015;**9:** e0004020.
9. Cameron DB, Mishra A, Brown AN. The growth of impact evaluation for international development: how much have we learned? *J Dev Effectiveness* 2016;**8:**1–21.

Commentary: Biases in the assessment of long-run effects of deworming

International Journal of Epidemiology, 2016, 2163–2165
doi: 10.1093/ije/dyw348
Advance Access Publication Date: 6 February 2017

Macartan Humphreys*

*Corresponding author. Columbia University and WZB. 812 IAB, $20 W 118th Street, New York, NY 10028.
E-mail: mh2245@columbia.edu

Accepted 6 September 2016

Introduction

Jullien and colleagues provide a critique of three working papers on the long-run effects of deworming interventions.[1] Despite being unpublished, these three papers have been prominent in the public debate in support of calls for such interventions over the past few years.[2] What can we really infer from them?

On first read, the critique by Jullien *et al*. is devastating. The three papers appear to have no redeeming qualities: a collection of fished results from poorly implemented and poorly analysed studies whose influence can only be explained by confirmation bias among deworming advocates.

On second read, and going back to the original papers, things are not so simple. A number of concerns described by Jullien *et al*. are on target. But a number seem to be off in ways that cannot be explained by differences in disciplinary norms.

I discuss the evaluation of this evidence according to possible sources of bias (mostly using Jullien *et al*.'s categories but adding some additional considerations).

Sources of Bias

Publicization bias

Consider first a type of publication bias. One might reasonably worry that these three publicized (but unpublished) studies, all displaying positive effects of deworming, were plucked by deworming advocates from a larger population of unpublished studies with many null or negative effects. However, although clearly it is hard to know where to look for unpublished (and unpublicized) null results, especially in the absence of preregistration norms, the fact that the search by Jullien *et al*. did not uncover any studies other than these three moderately increases confidence that the pattern of positive results is not simply a product of publicization bias.

Confounding bias

Jullien *et al*. worry about unknown bias due to absent baseline data in Baird *et al*.[3] For many social science experimentalists, this concern is hard to make sense of (at least if the assignment is considered to be as good as random), since unbiasedness is seen to stem from the assignment procedure, not the realization of assignments.[4] The concern with confounding in Ozier[5]—that observational variation is mixed up with experimental variation—also seems off. The key analysis provided in Ozier [Figure 1(B1)] clearly focuses on the experimental variation. Moreover, as the regression analysis includes fixed effects for cohorts, cohorts with no variation in treatment should effectively drop out. In both cases the economists could have made things easier by using a better randomization procedure and employing cleaner design-based inference procedures, but in neither case is there clear cause for concern.

Complex treatment bias

A second type of confounding bias relates to complex treatments. These are touched on only briefly by Jullien *et al.* but are possibly important, at least for Baird *et al.* and Ozier. In both cases, the treatment groups received more than medication and they had different exposure to research teams.

Spillover bias

Spillover effects can lead to underestimates of treatment effects or false-negatives, yet these risks seem not to concern Jullien *et al.* Paying more attention to these should if anything increase confidence in the positive findings.

Reporting and detection bias

Jullien *et al.* report high risks of bias due to non-blinded data analysis. To be clear, this reflects a disciplinary difference and there is no stated concern specific to these studies.

Attrition bias

Jullien *et al.* point to concerns about attrition bias that they labelled unclear risk in Ozier and high risk in Croke.[6] In both cases there were general concerns about out-migration. Out-migration can create a risk of bias, but there is no evidence of differential out-migration in either case. Moreover, it is hard to simultaneously believe that there are no long-run effects and that treatment induced major out-migration. The more study-specific concerns raised by Jullien *et al.* around attrition should not be considered sources of bias as they involve removal of non-experimental subjects (Ozier) or random sampling from experimental subjects (Croke): neither of these procedures generates bias, if done as described. Indeed, the removal of non-experimental subjects can remove a possible source of bias.

Selective reporting bias

Jullien *et al.* see large risks of selective reporting in Baird *et al.* The absence of a pre-analysis plan, the many varying versions of the paper and the many analyses make it hard to figure out what here is a test and what is exploration. In an earlier assessment, GiveWell[7] also examined the team's grant proposals as a type of analysis plan and noted substantial differences between this and the implemented analysis. The selective highlighting of some results, especially in the abstract, seems indisputable and supports the general concern that the overall conclusions taken from the analysis do not faithfully reflect the considerable variation in actual estimates.

There is one other less obvious dimension to the reporting that is not discussed by Jullien and colleagues. A confusing aspect of Baird *et al.* is that, although the authors describe the trial as comparing 50 schools that had 2-3 years of assigned free deworming with 25 that did not, the actual analysis compares 25 with 25. The other 25 schools had 2-3 years of assigned free deworming followed by a year of a cost-sharing treatment in 2001 and are, appropriately, effectively excluded from the main estimates. There is clearly confusion on this point, and when Jullien *et al.* focus on assignment and balance and so on they are working off the 50/25 comparison rather than the comparison that is actually used. If you conceptualize the analysis as a three-arm trial, this raises a question of why the reporting focuses on one pairwise comparison when others are implied by the design. The tables suggest for example that, though not reported, there is little evidence that the treatment '2 years free plus 1 year cost sharing' is effective relative to '1 year free.' As noted by Baird *et al.*, the fact that the estimated effects for the intermediate treatments are intermediate in magnitude is reassuring; but the fact that they are not themselves significant could raise worries.

Bias from subgroup analysis

Jullien *et al.* are critical of the subgroup analyses in Baird *et al.*, worrying that they reflect selective reporting. Baird *et al.* are however quite restrained on this front and focus almost uniquely on gender, which is standard in their field. It is almost unimaginable that analysis by gender would not have been in an analysis plan. Some other subgroup conditioning raises specific inference problems however: income source, for example, is a post-treatment category and conditioning on it can introduce bias. Clearly, pre-specification of analysis plans would have done a lot to remove many of these concerns.

Consider finally some sources of bias that get less attention, as follows.

Status quo bias?

Both sides are formally engaging in classical statistics and putting store by classical null hypothesis tests. Jullien *et al.* deem that 'an effect is present if $P < 0.05$' and dismiss estimates that do not reach this threshold. For example, the estimated effect (without controls) for mathematics scores in Croke is very large with an implied *P*-value that must be about 0.06. For Jullien *et al.* however, this means that the effect was not present. In the analysis of earnings in Baird *et al.*, a natural conditioning would focus on the older than school age sample, which is an exogenous category. When this is done, the estimated effects are very large—with gains around 20%—but power is lost and the *P*-value is 0.101, so this too is deemed not present. Surprisingly in other ways however, Jullien *et al.* do act like Bayesians.

They interpret null results as 'reasonable evidence of no effect', a view that is foreign to classical statistics.

The reliance on these sharp thresholds and complete discounting of evidence when *P*-values pass over an arbitrary threshold can create a bias towards inaction. It is especially questionable in situations when, as here, even modest effects estimated with great uncertainty could tip the cost/benefit calculation in favour of action for a rational decision maker.

Low power bias?

There is a lot of concern in Jullien *et al.* that the three studies are underpowered. It is hard to know exactly what Jullien *et al.*'s concern is here. *Ex ante* it might not have been wise to implement underpowered studies. But *ex post*, being underpowered does not alter the type I error rate nor render the *P*-values less credible. It does have two implications however. First, it should make us even more wary of treating no evidence for effects as evidence for no effects, as do Jullien et al. Second, it increases the risks of applying a 'statistical significance filter'[8], as Jullien et al do—that is, conditional on being significant, an estimate from an underpowered study is more likely to be too large in magnitude.

Failure to aggregate

These risks are compounded by a failure to aggregate inferences across studies. To see the problem here, consider the following idealized situation. You suspect a coin has a pro heads weighting. You flip it 20 times and it comes up heads 13 times. Your *P*-value for the null that the coin is fair is 0.26. You do two more independent trials. You get heads 14 times and then 12 times, yielding *P*-values of 0.12 and 0.50. Using the approach adopted by Jullien *et al.*, you conclude that all three trials support the view that the coin is fair. Yet coming up heads 39 times out of 60 tosses has an associated *P*-value of 0.03; and a Bayesian starting with a flat prior on the weighing of the coin would end up with a 99% posterior probability that the coin is weighted toward heads.

The obvious implication is that you should try to make inferences across studies and not apply the sharp thresholds study by study. Unfortunately, the deworming trials were not done in a way that makes such simple aggregation easy: outcomes are different, populations are different and two of the three trials are not independent in any case. The frustrating thing is that although such an ideal meta-analytic approach could not easily be used by Jullien *et al.*, this does not diminish the risks of the approach that is used.

Summary

In summary, it seems we currently have access to just three randomized controlled trial (RCT)-based studies examining long-run effects of deworming. These three studies contain many glimmers that point to long-term positive impacts of deworming. However, they come with risks. In all cases the RCTs were implemented before the emergence of norms around preregistration, specifically before the medical journals' statement on preregistration in 2005; and all were implemented by economists for whom blinding of analysts from treatment status is not common practice. So by contemporary medical standards, the research procedures used here fall short. There are also risks of inferential bias arising from complex treatments in two cases. However, many other risks considered by Jullien *et al.* are low or unclear, and a number of the specific criticisms raised by Jullien *et al.* do not stand up well. Moreover, applying a strict *P*-value threshold study by study, as do Jullien *et al.*, can produce a form of bias of its own. This all makes it hard to draw a neat conclusion. While scholars go back and forth, with one side showcasing every positive effect and the other every methodological flaw, policy decisions need to be made. There are two obvious, and not incompatible, ways forward. One is to implement the clean (and comparable and properly preregistered) trials to generate the kind of high quality evidence that will be needed to satisfy all sides. This may take many years. The other is to try to figure out a more Bayesian inferential strategy to aid decision making in the presence of imperfect evidence.

Conflict of interest: None declared.

References

1. Jullien S, Sinclair D, Garner P. The impact of mass deworming programmes on schooling and economic development: an appraisal of long-term studies. *Int J Epidemiol* 2016;**45**:2140–53.
2. Kremer M, Miguel E. *The Scientific Case for Deworming Children*. London: Thomson Reuters Foundation, 2015.
3. Baird S, Hicks JH, Kremer M, Miguel E. *Worms at Work: Long-run Impacts of a Child Health Investment*. Cambridge, MA: National Bureau of Economic Research, 2016.
4. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *J R Stat Soc A* 2008;**171**:481–502.
5. Ozier O. *Exploiting Externalities to Estimate the Long-term Effects of Early Childhood Deworming*. Washington, DC: World Bank Development Research Group, 2016.
6. Croke K. *The Long Run Effects of Early Childhood Deworming on Literacy and Numeracy: Evidence from Uganda*. Boston, MA: Harvard T. H. Chan School of Public Health, 2014.
7. GiveWell. Reanalysis of the Miguel and Kremer deworming experiment. 2012. http://www.givewell.org/international/technical/programs/deworming/reanalysis. (1 November 2016, date last accessed).
8. Gelman A, Weakliem D. Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist* 2009;**97**:310–16.