International Journal of Epidemiology, 2016, 2140–2153
doi: 10.1093/ije/dyw283
Advance Access Publication Date: 6 February 2017
Original article

Deworming

The impact of mass deworming programmes on schooling and economic development: an appraisal of long-term studies

Sophie Jullien, David Sinclair and Paul Garner*

Centre for Evidence Synthesis in Global Health, Liverpool School of Tropical Medicine, Liverpool, UK

*Corresponding author. Centre for Evidence Synthesis in Global Health, Department of Clinical Sciences, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK. E-mail: Paul.Garner@lstmed.ac.uk

Accepted 6 September 2016

Abstract

Background: Documents from advocacy and fund-raising organizations for child mass deworming programmes in low- and middle-income countries cite unpublished economic studies claiming long-term effects on health, schooling and economic development.

Methods: To summarize and appraise these studies, we searched for and included all long-term follow-up studies based on cluster-randomized trials included in a 2015 Cochrane review on deworming. We used Cochrane methods to assess risk of bias, and appraised the credibility of the main findings. Where necessary we contacted study authors for clarifications.

Results: We identified three studies (Baird 2016, Ozier 2016 and Croke 2014) evaluating effects more than 9 years after cluster-randomized trials in Kenya and Uganda. Baird and Croke evaluate short additional exposures to deworming programmes in settings where all children were dewormed multiple times. Ozier evaluates potential spin-off effects to infants living in areas with school-based deworming. None of the studies used pre-planned protocols nor blinded the analysis to treatment allocation.

Baird 2016 has been presented online in six iterations. The work is at high risk of reporting bias and selective reporting, and there are substantive changes between versions. The main cited effects on secondary school attendance and job sector allocation are from *post hoc* subgroup analyses, which the study was not powered to assess. The study did not find any evidence of effect on nutritional status, cognitive tests or school grades achieved, but these are not reported in the abstracts.

Ozier 2016 has been presented online in four iterations, without substantive differences between versions. Higher cognitive test scores were associated with deworming, but the appropriate analysis was underpowered to reliably detect these effects. The size of the stated effect seems inconsistent with the short and indirect nature of the exposure to deworming, and a causal pathway for this effect is unclear.

Croke 2014 uses a data set unrelated to the base trial, to report improvements in English and maths test scores. The analysis is at high risk of attrition bias, due to loss of clusters, and is substantially underpowered to assess these effects.

Conclusions: In the context of reliable epidemiological methods, all three studies are at risk of substantial methodological bias. They therefore help in generating hypotheses, but should not be considered to provide reliable evidence of effects.

Key words: Helminths, parasitic worms, children, cluster analyses, bias

Key Messages

- The long-term societal effects of mass deworming programmes for soil-transmitted helminths in low- and middle-income countries are contested.
- Advocates cite economic studies reporting long-term effects on health, schooling and economic development. We sought and appraised these studies using health technology assessment methods based on epidemiological principles.
- In the 11 reports from three studies, we found multiple potential sources of bias in the study methods, analysis and reporting. Of particular concern are: the lack of pre-planned protocols; multiple hypothesis testing followed by selective reporting of favourable results; and *post hoc* subgroup analyses.
- Our interpretation is that these trials do not provide credible evidence to support the claims of long-term effects. However, they raise interesting hypotheses that could be considered in further research.

Introduction

Soil- transmitted helminths remain common in many low- and middle-income countries, despite some evidence that global infection intensity may be declining.[1] The worms are unpleasant, cause discomfort and with heavy infections can undermine nutritional status and lead to serious complications.[2,3] It is obvious that children with symptomatic infection should be treated. It is also obvious that repeated mass treatment of whole communities with effective drugs will reduce the overall worm burden where helminths are common, at least in the short term.[4,5]

What is not obvious is whether mass deworming programmes have any measurable long-term effect on health and nutrition at the community level. A 2015 Cochrane review of trials administering multiple rounds of deworming treatment found little or no effect on average weight gain or average haemoglobin across 10 trials with more than 38 000 participants.[6] The review authors (which include one of the authors of this paper) interpreted this as reasonable evidence of no effect. Others have suggested that the trials were simply too short or were poorly designed for detecting effects once infected children are dispersed among large numbers of uninfected children.[7,8]

However, much of the advocacy and fund-raising for mass deworming programmes in children has drawn on studies reporting long-term effects on school attendance and economic development.[9–12] This advocacy contributed to the decision by India to run the largest national deworming programme in the world (targeting 270 million children in schools and preschools in 2016), and the Cochrane review has been criticized for excluding the studies cited for these effects.[10,13]

The objective of this paper was therefore to use health technology assessment methods based on epidemiological principles, to appraise the methods of these studies and to interpret their findings in the light of this appraisal.

Methods

Inclusion criteria

All follow-up studies based on randomized or quasi-randomized experimental trials (termed 'base trials') included in the 2015 Cochrane review. We included all outcome domains identified by the literature in this field as important for decision making and included in the Cochrane review (see Figure 1): nutritional status (measured by weight, height and haemoglobin); physical well-being (measured by exercise tolerance or self-reported measures); school attendance (measured by days present at school or years of school enrolment); and cognition and school performance (measured by formal tests and exam performance).[6] In addition, we included all economic productivity outcomes the author teams deemed to be reasonable overall measures.

Search strategy

We identified the main unpublished studies being cited, by reviewing the reference lists of prominent papers[9,13] and the webpages of deworming advocates.[10,11,14] We also

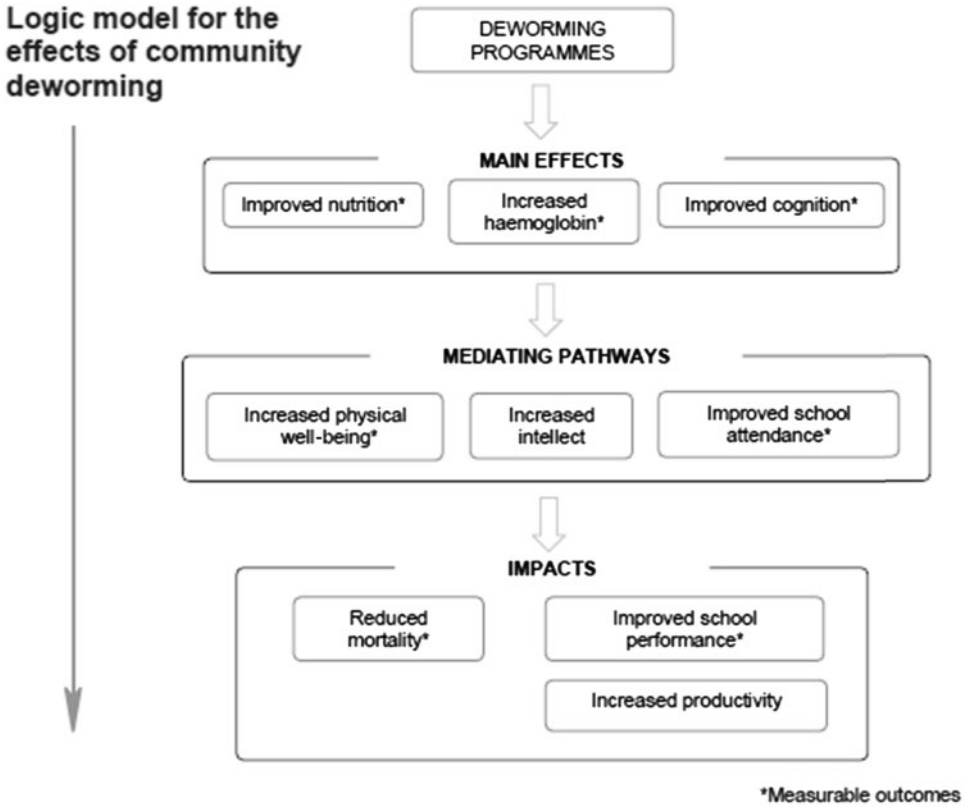

Figure 1. Logic model for the effects of community deworming.

Reproduced with permission from Taylor-Robinson 2015.[6]

searched Pubmed for published follow-up studies to all cluster-randomized studies included in the Cochrane review by using the search terms: ((deworm*[Title/Abstract] OR helminth*[Title/Abstract]) AND (Alderman[Author] OR Awasthi[Author] OR Hall[Author] OR Stoltzfus [Author] OR Wiria[Author] OR Rousham[Author] OR Miguel[Author])). Two authors independently screened the search results and applied the inclusion criteria.

Risk of bias assessment

For the base trials, we described the study design, setting, population, intervention and control. We assessed the risk of bias using the Cochrane tool for appraising randomized controlled trials and considered the potential for bias to also influence the results of the follow-up studies.[15]

For each follow-up study, we described the study design, population, timing, intervention and control group exposure to mass deworming, analytical approach, outcome measurement and reporting, and results. For the risk of bias assessment, we adapted the Cochrane tool for randomized controlled trials to take into account the additional risks posed by cross-sectional sampling from communities, many years after the planned experiment finished, as follows.

- Selection bias: we considered the randomization process of the base trial, the methods for selection of a proportional sample and the balance of potential confounders between groups.
- Measurement/ and detection bias: we considered the methods used to blind those collecting and analysing the data from the treatment allocation.
- Attrition bias: we considered loss of clusters, exclusion of participants after enrolment, migration in and out of the study area and the proportion and potential impact of missing outcome data.
- Selective reporting bias: we considered the use of a pre-planned protocol, the number of outcomes assessed and the potential for false-positive results, changes in the reporting of outcomes over time, and inclusion of important findings (showing association, or showing a lack of association) in the abstract.

For each domain, we classified studies as: 'low risk' when appropriate methods were described to reduce the potential for bias; 'high risk' when the methods described were inadequate to negate the potential for bias to influence the results; and 'unclear risk' when the impacts of any methodological problems were uncertain or there was insufficient information to make a clear judgment. We

refined these assessments after contacting the study authors for additional information.

Outcome credibility assessment

We first summarized the effect size and 95% confidence interval for all outcomes reported by the studies across the policy-important domains. As the included studies are outside the scope of what would normally be included in a Cochrane review, we familiarized ourselves with the study methods and findings and discussed which factors would be likely to be important when appraising the results. We then applied this appraisal systematically across studies.

We then further assessed the credibility of all the main findings reported in the abstracts, by considering: the evidence base for the stated effect from the main text (we considered an effect to be present if $P < 0.05$); the power of the study to detect this effect; the consistency of the effect across subgroups; the consistency of the effect across similar or related outcomes; and the robustness of the effect to adjustment for multiple inferences [although statistical adjustment for multiple testing is of limited value without a pre-planned analytical protocol, we considered the effect to be robust if the false discovery rate (FDR) q-value $<$ 0.05].[16] We also considered whether intermediate effects were present or absent on plausible causal pathways, and the plausibility of the effect in relation to the intensity of the intervention.

Results

We identified three unpublished, long-term follow-up studies[17–19] based on two cluster-randomized trials from Kenya and Uganda[20,21] (see Table 1). One additional study was excluded, as it was not based on a randomized experiment.[22]

All three follow-up studies are economic working papers available online but not formally published (Appendix 1, available as Supplementary data at *IJE* online). The study by Baird has been presented online in six iterations, although we were only able to access five (2011a, 2011b, 2012, 2015, 2016). The study by Ozier has been presented on-line in four iterations (2011, 2014, 2015, 2016). The search for published studies returned 94 on-line records, of which none were judged relevant to this review: 8 reports corresponded to the base trials in the Cochrane review, 11 were studies older than the base trials, 40 were not relevant to deworming interventions and the remaining 35 were not long-term follow-up studies.

Kenya trial (Miguel and Kremer 2004)

The base trial for the first two studies was conducted in Busia District, in Western Kenya, by Miguel and Kremer. The intervention comprised deworming drugs administered every 6 months, plus regular worm prevention education through public health lectures, wall charts and training of teachers. Seventy-five schools, with 32 565 pupils aged between 6 and 18 years, were allocated sequentially to one of three treatment groups: group 1 received the intervention from 1998, group 2 from 1999 and all groups received the intervention from 2001 onwards.

Risk of bias assessment. The quasi-randomized design means that there is a small risk of systematic differences between groups, and this risk will probably still be present in the follow-up studies. In addition, any effects observed in the follow-up studies may be attributable to the effects of the public health education activities rather than the anti-helminthic drugs. Although some would argue this is part of the intervention, it is not the main component of most large national deworming initiatives.[10] A complete risk of bias assessment is available in Table 2. Of note, an independent replication analysis of this trial was carried out in 2015, which found errors in the analysis of reported effects on haemoglobin and nutritional status; the authors now acknowledge that these effects are not 'statistically significant'. In a second replication that used the original authors analytical approach, the externalities were also not demonstrable, but the original trial authors have adjusted the parameters, conducted new analyses and contest this.[23–25]

Baird study (reported in a series of papers 2010-16)

The Baird series of reports presents analyses of a questionnaire survey of 5084 adults, 9 to 11 years after they participated in the Kenyan trial.[20] The analysis compares adults from schools which began receiving the intervention in 1998 and 1999, with adults from schools which did not receive the intervention until 2001 (see Appendix 2, available as Supplementary data at *IJE* online). As all participants eventually received the intervention, this study looks for effects attributable to the intervention group receiving an additional 2.4 years of the deworming intervention compared with the control group (Table 1). The paper presents data on nutritional, health, schooling and labour market outcomes.

Risk of bias assessment. The survey sampled adults from a complete list of all children who attended the schools in the base trial. The sample was selected using computerized randomization, and stratified by school, grade and gender (see Table 3). Baseline data were

Table 1. Characteristics of the base trials and the long-term follow-up studies

Study ID (versions)	Base trial					Follow-up study				
	Study ID	Country	Population	Number randomized (clusters)	Intervention	Population	Data collection	Sample size	Timing	Difference in deworming exposure between study groups
Baird series (2010, 2011a, 2011b, 2012, 2015, 2016)	Miguel and Kremer 2004	Kenya	School children aged between 6 and 18 years[a]	32 565[b] (75)	Deworming[c] every 6 months at school, plus health promotion	Adults aged 19 to 26 years who participated in the base trial as children	Questionnaire survey	5084	9 to 11 years after base trial started	2.4 additional years of deworming
Ozier series (2011, 2014, 2015, 2016)						Children aged 8 to 15 years who now attend the base trial schools, but were too young at the time of the trial to have participated	Field survey	21 309 for height and weight; 2371 for cognitive assessment	11 to 12 years after base trial started	Exposure to the 'spill-over' effects of deworming during the first year of life
Croke 2014 (2014)	Alderman 2006	Uganda	Pre-school children aged 1 to 7 years	27 995 (50)	Deworming[d] every 6 months at child health days (CHD)	Children aged 6 to 16 years who live in the area of the base trial and may have participated as children	Large-scale survey unrelated to base trial	763	10 to 11 years after base trial started	2 additional doses of deworming tablets

[a]In Miguel and Kremer 2004: girls aged 13 years or older were not intended to receive the drug intervention due to potential drug teratogenicity. However, some did receive deworming treatment.

[b]Miguel and Kremer 2004 was a quasi-randomized trial using sequential allocation.

[c]In Miguel and Kremer 2004, deworming medication was given as albendazole every 6 months (600 mg in 1998 and 400 mg in 1999) plus praziquantel at 40 mg/kg annually. It is estimated that 72% of children in the intervention and 5% of children in control groups received this.

[d]In Alderman 2006, deworming medication was given as albendazole 400 mg every 6 months. It is estimated that the deworming coverage increased from 21.7% before the intervention started in 2000 to 65.8% in 2003 in the intervention group, and from 23.9 to 34.6% in the control group (according to a cluster survey of households in all parishes, including 750 households in each group).

Table 2. Risk of bias assessments for the base trials

Study ID	Selection bias		Reporting and detection bias		Attrition bias	Other biases	
	Sample selection	Confounding	Blinding of outcome assessors	Blinding of data analysis		Contamination	Co-intervention
Miguel 2004	HIGH RISK • Systematic allocation (non-random) • Subsamples described as 'random' but no details given	UNCLEAR RISK • Groups broadly similar according to comparison of variables at baseline, but missing data to assess and confirm it	HIGH RISK • Not blinded	UNCLEAR RISK • Blinding not described	HIGH RISK • No clusters were lost • Considerable missing data for all outcomes.	LOW RISK • Deworming coverage of 5% in the control group • Transfer rate into a different school between 2% and 8%, with similar proportions among the three groups	HIGH RISK • Worm prevention education through regular public health lectures, wall charts and training of teachers[a] • Other school-based interventions simultaneously in 27/75 project schools
Alderman 2006	LOW RISK • Cluster randomized controlled trial	LOW RISK • Balanced baseline characteristics	HIGH RISK • Not blinded	HIGH RISK • Not blinded	LOW RISK • Two clusters were lost	HIGH RISK • Children dewormed in 2003: 65.8% in intervention group, 34.6% in the control group	LOW RISK • None

[a]Some may view this as part of the intervention, but current global policy advocates drug distribution, not intensive school health education.

presented for age and academic performance prior to the base trial, and although these appear balanced, this is probably insufficient to exclude the possibility of confounding due to the quasi-randomized design of the base trial. The analysis did not follow a pre-planned protocol, and those analysing the data were not blinded to treatment allocation.

The five versions available online to mid-2016 contain substantially different analyses which appear exploratory, and there is a high risk of false-positive results given the number of hypotheses tested for statistical significance increased from 228 in Baird 2011a to 650 in Baird 2016, largely due to the introduction of subgroup analyses (see Appendix 3, available as Supplementary data at *IJE* online). This process appears to be at high risk of reporting bias, and a narrative analysis suggests selective reporting as follows.

- Some outcomes reported in early versions were dropped from later versions. It is not made clear to the reader why, but it is likely to be due to the failure to demonstrate an effect (for example, cognitive test results reported in 2011a, but absent in 2016; with no apparent effect on Raven's matrices or English vocabulary).
- Effects are presented for outcomes which appear to be part of a larger undisclosed data set (for example, 'self-reported health rated as very good' presented without additional categories; and 'Kenyan women who participated as girls have fewer miscarriages' without presenting other health-related outcomes).
- Results from *post hoc* subgroup analyses are given prominence in the abstract and results (for example, an increase in secondary school attendance in females is claimed in the 2016 abstract, but no effect was apparent in the whole sample, and disaggregation by sex only appeared from 2012 onwards).
- The abstract changed substantially between versions, but none reported important findings of no effect (for example, there were no effects apparent on body mass index or height but these are not reported in any of the five abstracts; see Table 4 (and Appendix 4, available as Supplementary data at *IJE* online).

To further examine the influence of selective reporting, we compared the 'statistically significant' findings ($P < 0.05$) presented in the abstract, with the overall findings presented in Baird 2011a (Table 4). In the abstract, Baird reports that physical well-being, school enrolment and attendance and school performance or cognition are significantly higher in the group receiving earlier deworming. However, in the main text tables, only one of the seven outcomes measuring school performance/cognition is statistically significant. Similarly, for school attendance, an

Table 3. Risk of bias assessments of the long-term follow-up studies

Study ID	Selection bias		Reporting and detection bias		Attrition bias	Selective reporting
	Sample selection	Confounding	Blinding of outcome assessors	Blinding of data analysis		
Baird series	LOW RISK • Computer-generated random sampling from the eligible population, stratified by school, grade, and gender	UNCLEAR RISK • Age and academic performance before base trial appeared similar, but other potential confounders not presented • Uncertain risk of confounding due to the quasi-randomized design of the base trial	LOW RISK • Outcome assessors were unaware of how treatment would be defined in the analysis	HIGH RISK • Not blinded	LOW RISK • 2/75 clusters not included in the analysis • Effective tracking rate of 82.7%	HIGH RISK • No a priori analytical plan • Multiple significance testing • Inconsistency of outcome reporting over time • *Post-hoc* subgroup analyses presented as main results in the abstract • Important findings of no effect not reported in abstract
Ozier series	LOW RISK • Computer-generated random sampling from eligible population[a]	UNCLEAR RISK • Data on potential confounders are not provided separately for intervention and control groups • Only two cohorts (of seven) contain relevant randomized comparisons. Additional analyses of the whole sample are at uncertain risk of confounding due to secular trends[a]	LOW RISK • Outcome assessors were unaware of how treatment would be defined in the analysis	HIGH RISK • Not blinded	UNCLEAR RISK • Around 28% of sample excluded as they had migrated into the area after the base trial • Migration out of the area, which would represent missing data, is not well quantified • 2/75 clusters not included in the analysis	UNCLEAR RISK • No apriori analytical plan • Important finding of no effect on height not reported in abstract until the 2016 version. Data on weight not reported at all
Croke 2014	UNCLEAR RISK • Selection of villages described as 'random' but method not specified • Selection of households within villages by systematic selection	UNCLEAR RISK • Some confounders (access to water and private education) appear unbalanced	LOW RISK • Data were collected through a larger survey conducted for other reasons and unrelated to the base study	HIGH RISK • Not blinded	HIGH RISK • 28/50 clusters not included in analysis • Numeracy and literacy test outcomes available for 710/763 children (6.9% missing data) • Potential migration out of the area not addressed	UNCLEAR RISK • No a priori analytical plan

[a]Ozier series: of the seven annual cohorts, none of the children born in 1995 or 1996 lived in areas with active deworming programmes in the first year of life, whereas all the children born in 2001 did. Analyses across all seven cohorts therefore represent a mixture of randomized and observational data.

effect was only apparent in one of the three outcomes reported. Economic productivity was more complicated, as there were numerous subgroup analyses and a variety of derivative measures; an effect was apparent in 13 outcomes, with a further 19 reporting no statistically significant effect.

Credibility assessment. In Table 5 we attempt to provide a balanced presentation of the key results from Baird

Table 4. Assessment of selective reporting in Baird 2011a

Policy-important domains	Abstract		Tables and appendices	
	Number of outcomes reported as beneficial	Number of outcomes reported as no effect	Number of outcomes reported with $P < 0.05$	Number of outcomes reported with $P > 0.05$
Nutritional status	0	0	0	3
Physical well-being	1	0	2	0
School enrolment and attendance	1[a]	0	1	2
School performance and tests of cognition	1	0	1	6
Economic productivity	6[b]	0	13[c]	19[c]

[a]$P < 0.1$ and > 0.05.

[b]$P < 0.1$ and > 0.05.

[c]Economic productivity measured in: hours worked (seven subgroups); missed days (four subgroups); occupational subgroups (12); wage subsamples/derivative measures (nine).

2016, stratified by the policy-relevant outcome domains, and in Table 6 we present our credibility assessment for the outcomes reported in the abstract of Baird 2016.

In their 2016 abstract, Baird *et al*. state that men stayed 'enrolled for more years of primary school', and women were 'approximately one-quarter more likely to have attended secondary school'. These statements are supported by 'statistically significant' results within the text, but presentation of these two results in isolation could be regarded as misleading, as there is other information that is required for a balanced interpretation: (i) these effects were not present in the whole sample, and are only apparent in *post hoc* subgroup analyses which the analysis was not adequately powered to examine; (ii) neither result is robust to the authors' own adjustments for multiple inferences; and (iii) these are selected positive findings among a group of results for similar or related outcomes, that either show no effect (there was no evidence of an increase in the number of school grades attained in either sex), or provide an alternative explanation for these effects (those in the intervention group were actually more likely to have repeated a grade).

The abstract then uses these selected measures of educational effects to explain apparent shifts in the labour market, which are presented as beneficial. However, it is not clear to us which of these shifts represent a genuine economic improvement. For example, the number of hours women worked in agriculture appears lower in the intervention group and is presented as a benefit, but the number of hours worked by men appears higher. In reality, an effect in either direction could be interpreted as a benefit due to the alternative explanations of better health (enabling longer hours in manual work), or better education (enabling a move to higher skilled work). It is perhaps more useful to note that there was no evidence of an increase in hours worked in waged employment, and no evidence of an increase in non-agricultural earnings (waged earnings plus self-employed profits).

The authors clarified that the sample size was calculated to detect a 15% relative increase in secondary school attendance in the whole sample. The analysis was therefore not powered to look for subgroup effects. Furthermore, the sample size calculation does not seem to have been adjusted for the cluster design.

Ozier study (reported in a series of papers 2011-16)

The Ozier series report a field survey of 21 309 children attending the schools, quasi-randomized by the Kenyan trial 11 to 12 years earlier.[20] These children were too young at the time of the original trial to have received deworming treatment through the school-based programme. The analysis compares outcomes within each birth cohort from 1995 to 2001. Children aged less than 1 year living in communities where the deworming intervention had started are classified as the intervention group, and those living in communities where deworming had not yet started are classified as controls. The difference between these two groups is theoretically only that the children in the intervention group may have benefited from decreased worm prevalence among older siblings and the community during the first year of life, whereas the children in the control group did not.

Risk of bias assessment. The field survey conducted cognitive tests on a computer-generated random sample representing approximately 12 % of the eligible population. This sample covered seven annual school cohorts from 1995 to 2001. Only the 1998 and 1999 cohorts contain quasi-randomized comparisons relevant to the study question. In the 1995 and 1996 cohorts, none of the children lived in areas with active deworming programmes during the first year of life; and in 2001, all the children lived in

Table 5. Summary of effects reported in the long-term follow-up studies

Policy-important domains	Reported outcomes (unit of measurement)	Effect size (95% CI)		
		Baird series	Ozier series	Croke 2014
Nutritional status	Body mass index (kg/m^2)	0.02 kg/m^2 higher (0.07 lower to 0.11 higher)	-	-
	Height (cm)	0.11 cm shorter (0.65 shorter to 0.43 taller)	0.20 cm taller[a] (0.39 shorter to 0.80 taller)	-
	Haemoglobin (g/dl)	0.10 g/dl higher[b,c] (0.06 lower to 0.27 higher)	-	-
Physical well-being	Self-reported health status[d] (% rated as 'very good')	4.0% more (0.4 more to 7.6 more)	-	-
	Poor health in the past month (workdays missed)	0.11 days fewer[e] (0.38 fewer to 0.17 more)	-	-
School enrolment and attendance	School enrolment (%)	-	-	1.86% higher (0.72 lower to 4.44 higher)
	School enrolment (total years)	0.29 years more (0.00 more to 0.58 more)	-	-
	Secondary school attendance (%)	3.0% higher (4.0 lower to 10.0 higher)	-	-
School performance and tests of cognition	Had to repeat at least one grade (%)	6.3% higher (2.7 higher to 9.9 higher)	-	-
	Passed secondary school entrance exam (%)	5.0% higher (1.2 lower to 11.2 higher)	-	-
	Raven's matrices test score[f] (normalized scores, %)	1.1 % lower[g] (10.7 lower to 8.5 higher)	22.0% higher (6.4 higher to 37.6 higher)	-
	English vocabulary test score (normalized scores, %)	7.6 % higher[h] (3.4 lower to 18.6 higher)	16.1% higher (3.1 lower to 35.3 higher)	16.4% higher (17.74 lower to 50.54 higher)
	Math score (normalized scores, %)	-	-	301 % higher (0.81 lower to 61.0 higher)
Economic productivity	Hours worked per week (hours)	1.58 h more (0.50 fewer to 3.66 more)	-	-
	Monthly earnings (waged employment plus self-employed earnings)	226 higher[i] (1162 lower to 1614 higher)		
	Monthly earnings (waged employment only)	26.9% more (9.9% more to 43.9% more)	-	-

[a]Ozier also reports height-for-age and stunting, which are consistent with the findings for height.

[b]Baird 2011a reported control group estimate of 126.1 and coefficient estimate of 1.03 but no unit of measure, and we asume they used g/l (SI units); we report this outcome as g/dl.

[c]Findings on haemoglobin are not reported in the Baird 2016 version, but are in Baird 2011a and 2011b.

[d]The Baird series also report the proportion of women who had experienced a miscarriage, which was lower in the intervention group. It is excluded from this table as it seems a spurious outcome to present in isolation without measuring a large range of other potential health outcomes.

[e]Findings on workdays missed due to poor health in the past month are not reported in the Baird 2016 version, but are in Baird 2011a. In Baird 2011b, this outcome is reported for the out-of-school subsample only.

[f]Ozier used the 12 questions in set B of the Raven's Progressive Matrices. Baird gives no further details on the questions used for assessing the Raven's matrice test score.

[g]Findings on Raven's matrices test score are not reported in the Baird 2016 version; they are in Baird 2011a only.

[h]Findings on English vocabulary test score are not reported in the Baird 2016 version, but are in Baird 2011a, 2011b and 2012.

[i]The unit of this outcome is not reported, although we could assume it is the local currency.

Table 6. Outcome appraisal of all outcomes reported in the abstract of Baird 2016

Outcomes reported in the abstract		Evidence base for stated effect	Effect present in whole sample?[a]	Effect robust to adjustment for multiple inference?[b]	Effect consistent across related outcomes?[c]	
Men	'Stay enrolled for more years of primary school'	Men from intervention areas had higher total years enrolled in primary school ($P < 0.05$)	Yes	No	No	No statistically significant difference in the total number of school grades attained ($P > 0.1$), and adults from intervention areas more likely to have repeated at least one grade ($P < 0.01$)
	'Work 17% more hours each week'	Men from intervention areas worked more hours in the past week ($P < 0.05$)	No	No	–	–
	'Spend more time in non-agricultural self-employment'	A borderline effect on hours worked in non-agricultural self-employment in men ($P < 0.1$)	Yes ($P < 0.05$)	Remains borderline	No	No statistically significant difference in monthly non-agricultural earnings ($P > 0.1$)
	'Spend more time in manufacturing'	Men from intervention areas had a higher manufacturing job indicator ($P < 0.05$)	Yes	No	No	No statistically significant effect on hours worked in waged employment ($P > 0.1$), and no statistically significant difference in monthly non-agricultural earnings ($P > 0.1$)
	'Miss one fewer meals per week'	Men from intervention areas had eaten more meals the previous day ($P < 0.01$)	Yes	Yes	–	–
Women	'One-quarter more likely to have attended secondary school'	Women from intervention areas had higher secondary school attendance ($P < 0.05$)	No	No	No	No statistically significant difference in the number of school grades attained ($P > 0.1$)
	'Reallocate time from traditional agriculture into cash crops'	Women from intervention areas had a higher 'grows cash crop' indicator ($P < 0.05$)	Yes	No	–	–
	'Reallocate time from traditional agriculture into non-agricultural self-employment'	Women from intervention areas worked more hours in non-agricultural self-employment in the past week ($P < 0.05$)	Yes	No	No	No statistically significant difference in monthly non-agricultural earnings ($P > 0.1$)

[a]The subgroup analysis by sex was not introduced until the third edition of the Baird series and so is considered *post hoc*. We considered the effect to be present in the whole sample if $P < 0.05$ for both sexes combined.

[b]The authors of the Baird series conducted adjustments for multiple inference. We considered the effect robust to adjustment if the FDA q-value < 0.05.

[c]With so many outcomes presented, we considered whether the effects of related outcomes consistently suggested benefit.

areas with active deworming programmes. Analyses across the whole sample (seven cohorts) are thus secondary observational analyses, with unknown secular changes potentially confounding the findings (see Table 3). Data collection was appropriately blinded to treatment allocation, but again data analysis was not blinded and was not guided by a pre-planned protocol. Important findings of no apparent effect on height and height-for-age were not

reported in the abstract until the 2016 version (despite being one of the main a priori hypotheses, according to communication with the authors). Although weight data were collected for 21 309 children, they were not part of the analysis and not presented.

Credibility assessment. In the 2016 abstract, Ozier states that exposure to the spill-over effects of deworming programmes during the first year of life produced 'large cognitive effects, comparable to between 0.5 and 0.8 years of schooling'. This statement is based on demonstrable effects on two out of five cognitive tests (Raven's matrices and verbal fluency; $P < 0.05$) and a trend towards benefit on all five tests. These positive effects are taken from analyses across the whole sample, which include non-randomized data. However, following communication with the authors, additional tables were produced confirming these effects were still apparent in analyses limited to the quasi-randomized cohorts from 1998 and 1999. It should, however, be noted that the revised analysis is substantially underpowered to reliably detect these effects (communication with the authors confirmed that only the analyses including all seven annual cohorts were adequately powered). The authors themselves explain the lack of effect on height to be related to the low worm load in young children. We consider this observation, along with the very low intensity of the intervention being tested, to question the plausibility of the stated effect.

Uganda trial (Alderman 2006)

The base trial for the third study[18] was conducted in Eastern Uganda by Alderman *et al.*[21] The intervention was implemented through Child Health Days (CHD) and comprised albendazole 400 mg every 6 months. Fifty parishes in five districts were identified as having heavy worm loads and randomly allocated to the intervention and control arms. Over the 3-year programme from 2000 to 2003, children in both groups attended 1.74 CHDs on average, with only the intervention group scheduled to be dewormed but both groups receiving additional health services such as vaccination and health promotion. Participants were pre-school children aged between 1 and 7 years, and deworming became routine and free for everyone shortly after the end of the study.

Risk of bias assessment. The base trial used a truly random method of allocation (a coin toss), but although deworming was the intended difference between intervention and control groups, up to 35% of those in control areas were also dewormed, from private clinics or shops (see Table 2).

Croke study 2014

Croke uses a large-scale questionnaire survey conducted in Uganda 7 to 8 years after the end of the trial by Alderman.[21] The survey was unrelated to the base trial but covered some of the same parishes, and included 763 children who would have been aged between 1 and 7 years at the time of the base trial and who therefore might have participated. The study compares children living within the intervention parishes of the base trial with children living in the control parishes. The difference between the two groups (ignoring migration in and out of the area) is therefore likely to be less than two additional doses of albendazole during the 3 years of the programme. The analysis reports on numeracy and literacy test outcomes.

Risk of bias assessment. The sampling method is reported as random, but the descriptions of sampling are inadequate to make a clear assessment of the risk of selection bias (see Table 3). Data acquired through correspondence with the author reports on 11 covariates, among which the treatment group appears to have had better access to water (24% of individuals compared with 3%) and private education (14% compared with 9%). The data collection process was unrelated to the deworming base trial and so unlikely to have been influenced by it, but data analysis was not blinded. The risk of attrition bias is high with only 22 of the 50 parishes recruited by Alderman included in the sample (10 from the intervention group and 12 from the control group), and no assessment of the effects of migration. There was no pre-planned protocol.

Credibility assessment. Croke states that children who lived in intervention parishes during the base trial period had 'test scores 0.2 to 0.4 standard deviations higher than those in control parishes'. Setting statistical significance at $P < 0.05$, this effect was not present in the raw data and only apparent after adjustment for age, gender and survey year. No formal power calculations were conducted and the analysis is substantially underpowered to detect these effects, with less than a third of the sample size calculated by Ozier. The authors found no evidence of an effect on school enrolment, but do not report this in the abstract.

Discussion

In summary, of the three included long-term follow-up studies: the Baird series reports possible effects on secondary school attendance and job sector choices, 9 to 11 years after a head start of 2.4 years of additional school-based deworming; the Ozier series reports possible externalities on cognitive development in children living in areas with school-based deworming during the first year of life; and Croke 2014 reports possible effects on English and maths test scores 10 to 11 years after less than two additional

doses of deworming tablets during early childhood. All the reports present these as clear evidence of benefit of deworming programmes.

Long-term studies of the effects of public health interventions are complex and difficult to do. We therefore acknowledge the hard work of the study authors and research teams. However, from our epidemiological standpoint we find substantial reason to doubt the validity and plausibility of these findings, given the information provided and the process of analysis that has been documented. As such, we believe they should be regarded as hypothesis-generating, rather than as reliable evidence of effects to support large-scale deworming programmes in low- and middle-income countries.

First, we note that these studies do not provide the evidence of cumulative effects from multiple rounds of deworming, that some have called for and others have attributed to them.[8,26] In all three studies, most participants in both the intervention and control groups would have been 'dewormed' multiple times during their preschool or school years, and the largest 'intervention' under evaluation was an additional 2.4 years of deworming medication in the Baird series. The majority of children in these studies would therefore be worm-free, or would have reduced worm counts during much of their childhood, and consequently the consistent finding of no effect on height or weight across all three studies is unsurprising. More subtle nutritional pathways for the observed effects, such as via micronutrient status, also seem unlikely to act over such short durations.

Second, we are concerned about the selective reporting of favourable results in the abstracts, especially after multiple significance testing and *post hoc* subgroup analyses. All three papers herald from an economic discipline, but we assess them against current epidemiological standards and make no apology for that. The policy under evaluation is a public health programme, and the potential for bias exists irrespective of discipline. We do however acknowledge that some of the problems exist, at least in part, due to current norms within economics and the reporting requirements of economic journals (such as strict word limits for abstracts). Whereas some economists may argue that the accuracy of conclusions is improved over time through the refinement and addition of new analyses, we are concerned that the process risks cumulative selective reporting, and our analysis provides some indicators that this may be the case.

Of note, none of these three studies worked to a preplanned analytical protocol, and although this has been a standard requirement within epidemiology for some time, it has only recently been recognized as important within economics.[27] This was also noted in the replication analyses of the primary trials.[23,24] Nevertheless, this approach may well produce misleading results and conclusions, and statistical correction of multiple testing is insufficient to correct selective reporting.

The abstract to the Baird series exclusively presents the positive results, and leaves readers unaware of the multiple findings of no evidence of effect and the conflicting findings within the analysis. For example:

- no evidence of effect on markers of nutrition (weight or height);
- no evidence of effect on multiple tests of cognition (the same tests as reported by the Ozier series);
- an increase in the need to repeat a school grade in the intervention group (an alternative explanation for the observed increase in years in school, and consistent with the finding of no overall increase in the number of grades achieved);
- no evidence of effect on secondary school attendance prior to subgrouping by gender (only the whole sample is adequately powered to detect an effect), suggesting a potentially spurious subgroup finding;
- little evidence of effect on secondary school attendance in females after adjustment for multiple significance testing ($P = 0.084$ after adjustment);
- no evidence of effect on monthly earnings.

We do know that *post hoc* analyses increase the risk of type 1 errors (finding an effect when there is no effect present).[28,29] Item 18 of the 2010 CONSORT statement for the reporting of randomized trials specifies that *post hoc* analyses should be clearly labelled as such and considered as exploratory. In addition, the explanatory note states that '*Post hoc* subgroup comparisons (analyses done after looking at the data) are especially likely not to be confirmed by further studies'.[30]

At face value, there is a consistency of findings across the two remaining studies by Ozier and Croke. Both studies have substantial methodological and plausibility limitations which should temper their interpretation, but the observed effect after such a small deworming exposure probably deserves further consideration and should be amenable to testing through well-designed randomized trials.

More generally, there appears to be a tendency for advocates of deworming to 'build a case' for deworming, by drawing together evidence which supports their prior beliefs and ignoring or dismissing the evidence that does not.[9,31,32] This 'confirmation bias' is common, but runs counter to current standards in transparent, evidence-informed decision making[33] and has led to the claims of these studies being cited verbatim without appropriate appraisal.[34]

Government ministries responsible for resource allocation; philanthropists supporting these programmes, and the public who are subjected to them, require transparency about what effects could reasonably be expected. If a

community in a given setting has a high prevalence of untreated worm infections, then mass deworming programmes may well be an effective way to reach and treat a large number of children. If however, the problem is poor school attendance or low educational attainment, then these are problems which probably require different solutions.[35]

Conclusion

These three studies all have substantial problems in their methods and analysis, which leave unanswered questions about the use of these studies to justify the effectiveness of deworming programmes. They help in generating hypotheses. Decisions about whether or not to implement mass treatment programmes, calculations around programme cost and advocacy to the public, should be based on reliable estimates of effects, informed by robust evidence.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

This work was supported through the Effective Health Care Research Programme Consortium, funded by UK aid from the UK Government for the benefit of developing countries (Grant: 5242). The views expressed in this review do not necessarily reflect UK government policy. The funding source had no role in identifying the research topic, nor in the design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of interest: P.G. is Director of the Evidence Building and Synthesis Research Consortium that receives money to increase the number of evidence-informed decisions by intermediary organizations, including WHO and national decision makers, which benefit the poor in middle- and low-income countries. D.S. and S.J. are employed as part of this Consortium. P.G. is the co-ordinator of a WHO Collaborating Centre for Evidence Synthesis for Infectious and Tropical Diseases [http://apps.who.int/whocc/default.aspx; UNK234] and one of the Centre's aims is to help WHO in its role as an intermediary in communicating reliable summaries of research evidence to policy makers. P.G. is an author of the Cochrane review evaluating the effects of community-based deworming on health, nutrition and school participation. P.G. receives support from COUNTDOWN, a grant to the Liverpool School of Tropical Medicine from the Department for International Development to promote control of neglected tropical diseases in developing countries, including soil-transmitted helminths.

References

1. de Silva NR, Brooker S, Hotez PJ, Montresor A, Engels D, Savioli L. Soil-transmitted helminth infections: updating the global picture. *Trends Parasitol* 2003;**19**(12):547–51.
2. Bethony J, Brooker S, Albonico M *et al*. Soil-transmitted helminth infections: ascariasis, trichuriasis, and hookworm. *Lancet* 2006;**367:**1521–32.
3. Baba AA, Ahmad SM, Sheikh KA. Intestinal ascariasis: the commonest cause of bowel obstruction in children at a tertiary care center in Kashmir. *Pediatr Surg Int* 2009;**25:**1099–102.
4. WHO. *Prevention and Control of Schistosomiasis and Soil-transmitted Helminthiasis*. Report of a WHO Expert Committee, Technical Report Series, No. 912. Geneva: World Health Organization, 2002.
5. Mwandawiro Charles. *A New Perspective on the War on Worms*. 2015. http://www.impatientoptimists.org/Posts/2015/09/A-New-Perspective-on-the-War-on-Worms#.Vukw52SLTZu (15 March 2016, date last accessed).
6. Taylor-Robinson D, Maayan N, Soares-Weiser K, Donegan S, Garner P. Deworming drugs for soil-transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin, and school performance. *Cochrane Database Syst Rev* 2015;**7**:CD000371.
7. Walson JL. Don't shoot the messenger. *PLoS Negl Trop Dis* 2015;**9:**e0004166.
8. Montresor A, Addiss D, Albonico M *et al*. Methodological bias can lead the Cochrane Collaboration to irrelevance in public health decision-making. *PLoS Negl Trop Dis* 2015;**9:**e0004165.
9. Ahuja A, Baird S, Hicks JH, Kremer M, Miguel E, Powers S. When should governments subsidize health? The case of mass deworming. *World Bank Econ Rev*, 2015. doi:10.1093/wber/lhv008.
10. Evidence Action. *Deworm the World Initiative*. 2007. http://www.evidenceaction.org/dewormtheworld (3 March 2016, date last accessed).
11. Abdul Latif Jameel Poverty Action Lab. *Deworming: A Best Buy for Development*. J-PAL Policy Bulletin, 2012. https://www.povertyactionlab.org/sites/default/files/publications/2012.3.22-Deworming.pdf (3 March 2016, date last accessed).
12. WHO. *Deworming Campaign Improves Child Health, School Attendance in Rwanda*. 2015. http://www.who.int/features/2015/rwanda-deworming-campaign/en/ (3 March 2016, date last accessed).
13. Hicks JH, Kremer M, Miguel E. The case for mass treatment of intestinal helminths in endemic areas. *PLoS Negl Trop Dis* 2015;**9:**e0004214.
14. Evidence Action. Give Well. Deworm the World Initiative. 2015. http://www.givewell.org/international/top-charities/deworm-world-initiative (14 March 2016, date last accessed).
15. Higgins JP, Altman DG, Sterne JA. Assessing risk of bias in included studies. In: Higgins JP, Green S (eds). *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0*. The Cochrane Collaboration, 2011. http://handbook.cochrane.org/ (7 March 2016, date last accessed).
16. Noble WS. How does multiple testing correction work? *Nat Biotechnol* 2009;**27:**1135–37.
17. Baird S, Hicks JH, Kremer M, Miguel E. *Worms at Work: Long-Run Impacts of a Child Health Investment*. 2016. http://www.nber.org/papers/w21428 (12 January 2016, date last accessed).
18. Croke K. *The Long Run Effects of Early Childhood Deworming on Literacy and Numeracy: Evidence from Uganda*. 2014. http://scholar.harvard.edu/files/kcroke/files/ug_lr_deworming_071714.pdf (18 August 2016, date last accessed).
19. Ozier O. *Exploiting Externalities to Estimate the Long-Term Effects of Early Childhood Deworming*. 2016. http://economics.ozier.com/owen/papers/ozier_early_deworming_20160727.pdf (18 August 2016, date last accessed).

20. Miguel E, Kremer M. Worms :identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 2004;**72:**159–217.
21. Alderman H, Konde-Lule J, Sebuliba I, Bundy D, Hall A. Effect on weight gain of routinely giving albendazole to preschool children during child health days in Uganda: cluster randomized controlled trial. *BMJ* 2006;**133:**122.
22. Bleakley H. Disease and development: evidence from hookworm eradication in the American South. *Q J Econ* 2007;**122:**73–117.
23. Aitken AM, Davey C, Hargreaves JR, Hayes RJ. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication. *Int J Epidemiol* 2015;**44:**1572–80.
24. Davey C, Aitken AM, Hayes RJ, Hargreaves JR. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped wedge trial. *Int J Epidemiol* 2015;**44:**1581–92.
25. Hicks JH, Kremer M, Miguel E. Commentary: Deworming externalities and schooling impacts in Kenya: a comment on Aitken *et al.* (2015) and Davey *et al.* (2015*). Int J Epidemiol* 2015;**44:**1593–96.
26. Medley GF, Hollingsworth TD. MDA helminth control: more questions than answers. *Lancet Glob Health* 2015;**3:**e583–84.
27. Miguel E, Camerer C, Casey K *et al*. Promoting transparency in social science research. *Science* 2014;**343:**30–31.
28. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995;**310:**170.
29. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess* 2001;**5**(33):1–56.
30. Moher D, Hopewell S, Schulz KF *et al.* CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;**340:**c869.
31. World Health Organization. *The Evidence is in: Deworming Helps Meet the Millennium Development Goals*. Geneva: WHO, 2005.
32. Albonico M, Allen H, Chitsulo L, Engels D, Gabrielli A-F, Savioli L. Controlling soil-transmitted helminthiasis in pre-school-age children through preventive chemotherapy. *PLoS Negl Trop Dis* 2008;**2**(3):e126.
33. Oxman AD, Lavis JN, Lewin S, Fretheim A. SUPPORT Tools for evidence-informed health Policymaking (STP) I: What is evidence-informed policymaking? *Health Res Policy Syst* 2009;**7(Suppl 1):**S1.
34. Lord CG, Ross L, Lepper MR. Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. *J Pers Soc Psychol* 1979;**37**(11):1098–109.
35. Krishnaratne S, White H, Carpenter E. *Quality Education for all Children? What Works in Education in Developing Countries*. Working Paper 20. New Delhi: International Initiative for Impact Evaluation (3ie), 2013.

International Journal of Epidemiology, 2016, 2153–2156
doi: 10.1093/ije/dyw350
Advance Access Publication Date: 6 February 2017

Commentary: Assessing long-run deworming impacts on education and economic outcomes: a comment on Jullien, Sinclair and Garner (2016)

Sarah Baird,[1] Joan Hamory Hicks,[2] Michael Kremer[3] and Edward Miguel[4]*

[1]Milken Institute School of Public Health, George Washington University, Washington, DC, USA, [2]University of California, Center for Effective Global Action, Berkeley, CA, USA, [3]Department of Economics, Harvard University and NBER, Cambridge, MA, USA and [4]Department of Economics, University of California, Berkeley, and NBER, Berkeley, CA, USA

*Corresponding author. Department of Economics, 508-1 Evans Hall #3880, University of California, Berkeley, CA 94720, USA. E-mail: emiguel@berkeley.edu

Accepted 6 September 2016

Introduction

Jullien, Sinclair and Garner (2016)[1] (henceforth JSG) state that they seek to 'appraise the methods' of three recent papers that estimate long-run impacts of mass deworming on educational or economic outcomes. This commentary focuses on their discussion of Baird, Hicks, Kremer and Miguel (2016)[2] (henceforth Baird). We welcome scrutiny of our work, and appreciate the opportunity to discuss JSG.[1]