Summary of Evaluation of Kremer et al. (2021)

Prepared by Megan D. Higgs, Ph.D. Statistics, Critical Inference LLC

Updated 10/4/2021

Contents

1 Introduction

Kremer et al. (2021) reports on a meta-analysis used to combine data on child (under age 5) mortality from 10-13 randomized controlled trials (RCTs) that investigated water treatment interventions, with a focus on chlorination. The goal of the meta-analysis was to obtain an overall estimate of the expected decrease in the probability of mortality for children under 5 due to water chlorination. The studies ultimately included in the analyses are heterogeneous in terms of location, time frame, ages included, actual interventions, and lengths of time children were monitored. The wording in the manuscript implies broad generalizability of the results in terms of a reduction in under-5 mortality to other locations and times; as well as recommendations for directly using the resulting estimates in calculations leading to statements about potential lives saved per cost of interventions.

1.1 Overview of this evaluation and summary

This document summarizes an evaluation of the methods, assumptions, and interpretations presented in Kremer et al. (2021) based on my expertise as statistician. My evaluation is based on the version of the Kremer

et al. (2021) manuscript and supplemental materials provided to me by GiveWell, and other information provided by GiveWell about how the estimate may ultimately be used; it does not include evaluation of the individual studies used in the meta-analysis. This summary is meant to provide a high level overview of primary concerns and recommendations, and does not provide detailed technical concerns related to statistical analyses.

In general, meta-analysis provides an attractive endpoint in terms of synthesizing information from multiple studies into a single quantitative summary with greater precision than those obtained from individual studies. However, the challenges and limitations of meta-analysis are under-appreciated, often leading to overconfidence in the results. As with any analysis, results are dependent on the studies included - including designs and data - and assumptions used in the statistical modeling, all of which affect justification for generalizations ultimately made. It is common for reports of meta-analyses to over-generalize conclusions relative to what is actually justified directly from the studies included and modeling used.

1.2 Overall summary of conclusions

Overall, the report provided by Kremer et al. (2021) falls into the common trap of over-generalizing results relative to the data and assumptions going into the analyses. Specifically, wording implies a broad interpretation of the main quantitative result as an estimate of the reduction in under-5 mortality due to water treatment for any region and time frame; while limitations of, and heterogeneity in, the studies used do not directly support such statements and generalizations. Additionally, transparency in reporting the research process, particularly the exclusion process leading to the final 10-13 studies included in the main analyses, could be more clearly described to improve transparency and reproducibility of methods, which would in turn help justify the extent to which results may be generalized.

From a statistical perspective, the estimated reduction in the probability of mortality from the meta-analysis should be taken as one source of information about the potential effect of chlorination interventions as summarized through the combinations of results from the specific 10 RCTs included, acknowledging the differences in designs and interventions. The uncertainty expressed in confidence/credible intervals is likely substantially under-stated given the heterogeneity among the included studies, especially in the time period over which children were monitored post-intervention. Also, the meaning of the quantity represented by the overall point estimate should be carefully considered relative to how the outcomes were obtained for the individual studies - a jump to "under-5 mortality" may be arguable, but is in need of additional information, assumptions, and justification.

2 Primary concerns

This summary addresses two general areas that should be taken into account when weighing how to use the results and conclusions to inform decisions regarding allocation of resources to water chlorination efforts: (1) general transparency and reproducibility of the work, and (2) generalization of the results, including meaning of the combined estimate, careful scrutiny of inclusion/exclusion criteria, and assessing stated levels of uncertainty.

2.1 Transparency, reproducibility, and justification

2.1.1 Inclusion/exclusion criteria

For any meta-analysis, a crucial part of the process is very clear documentation of the path leading to the final collection of studies and data used for the analysis. The results and potential for their generalization depend on the set of studies ultimately included, as well as those excluded. Selection of studies is generally open to many researcher degrees of freedom and therefore transparent presentation of the decision tree of exclusion criteria, as well as justification for those criteria should be an integral part of the analysis, and the process should be presented in a way that it is reproducible by others with access to the same materials. Additionally, results can be strengthened and placed in context by in-depth discussion of how the studies

ultimately included may differ from those excluded (inclusion/exclusion bias); such an exercise addresses what the results might reasonably represent.

The Kremer et al. (2021) manuscript could benefit from greater clarity and transparency in presentation of the inclusion/exclusion process. In the abstract, the authors state that they combined "all existing RCT evidence on mortality," but it is possible that some smaller or older studies were not included and care should be taken in not overstating the inclusion criteria relative to inherent limitations in identifying studies. In one place the process is described as "systematically identified all RCTS in developing countries examining the impact of water interventions on child diarrhea from 1970 to 2020 ... The search resulted in thirteen studies." This process of searching for RCTs and finding a total of 13 is inherently different from the process ultimately described in the Methods section and in Figure 3 that depicts a workflow starting from 1411 studies identified via a database search plus 73 from Wolf et al., and then excluding studies based on a strict screening process. Exclusion criteria and steps used to apply them are not well described. Changes in presentation and wording could better convey the details of how the process was actually carried out, even if that process differs from expectations for how a pre-registered meta-analysis would be carried out. Implications of the exclusion/inclusion criteria on ultimate inferences are discussed following sections.

2.1.2 Statistical analyses

Meta-analysis also requires many decisions about statistical methods and associated assumptions. The Kremer et al. (2021) manuscript does provide very brief descriptions of the methods used, but the justification of choice of methods and addressing of key assumptions is superficial. For example, a method being considered "standard" is not adequate justification for its use in a particular instance. Additionally, the models used are not explicitly provided, references used to back up the choices could be improved, and software used for modeling and creation of graphics should all be explicitly described and referenced. As currently presented, it would be difficult to reproduce the analyses; additional information would be needed, including raw data used and computer code. The raw data used in the analyses (counts of mortalities and non-mortalities for the treatment and control groups from the different studies) should be included in the paper, even if only in graphical form. Going forward, the researchers might consider collaborating with a statistician with expertise in logistic regression, survival analysis, and Bayesian modeling.

2.2 Cautions related to generalizing the results

A challenging part of any statistical analysis is justification of the scope of inferences ultimately made - based on careful consideration of the study design, data, analyses, and any additional assumptions or arguments. Meta-analyses can prove particularly challenging in this regard, as the motivation for carrying out the analysis is often the hope for broad generalizability of results; however, appropriate inferences are still subject to limitations dependent on the information going into the analysis. The act of combining information from multiple studies does not imply the results apply generally without constraint, and generalizations should be explicitly acknowledged and justified. Several concerns relative to generalizations made in Kremer et al. (2021) are discussed in this section.

2.2.1 Discussing potential limitations based on studies included

The Kremer et al. (2021) report does not explicitly discuss the limitations of what the 10-13 studies included represent relative to the over 1400 studies originally identified, or limitations in how the results should be used by the authors or others in different contexts in the future. A first step in deciding how to generalize results of a meta-analysis is an in-depth discussion of how the small subset of studies ultimately used in the analysis may differ from the much larger collection of those not used in the analysis, from speculating on those never identified to specifically addressing the 51 studies for which the authors were contacted, but data were not obtained for various reasons. The authors can at least speculate on differences between the 10 ultimately used and the larger collections, and how those differences or "what's left out" might affect inferences. This point then leads to questioning of statements made in the manuscript and is related to the earlier concern regarding lack of clarity or transparency in the inclusion/exclusion process - "all existing RCT evidence" is not the same as "all RCT evidence we could find based on the effort we could afford."

Beyond how the studies included may differ from those excluded, is expectation for more in-depth discussion of the differences among the included studies. Another key part of a meta-analysis is justifying that it makes sense to combine the results of different studies into a single quantitative summary - assuming the results of the studies represent estimates of the same underlying quantity or that an overall mean captures the quantity of interest. This is difficult to judge, but can be explicitly discussed and justified based on careful consideration of study differences and similarities. The same principles that would apply to sampling individuals from a population apply here; for example, if we were going to use information from only 10 individuals to describe a very diverse population of people, we would have to provide details about how the 10 people were selected and why their combined information should be taken to represent the broader population. What assumptions would be necessary and how reasonable do we think those assumptions are? It may be ultimately be deemed reasonable to make such general statements found in Kremer et al. (2021), but arguments and explanations that would be needed for the justification are not currently addressed in the manuscript.

Kremer et al. (2021) do include a comparison of the diarrhea prevalence of the 13 studies to the 2017 estimates from lower and middle-income countries. This type of argument and discussion is useful if carried further, and actual data points can be plotted instead of using a histogram and only the minimum, maximum, and weighted average. The usefulness of the Kremer et al. (2021) work could be greatly improved by including more of this within the manuscript, to help readers understand how to ultimately interpret and judge the results.

2.2.2 Combining estimates across studies

When combining information across studies that differ in their experimental designs, it is important to ask what the overall combined estimate actually represents. This question can be easy to overlook in meta-analyses, as it is common to proceed under the implicit assumption that a single overall effect is meaningful and needed, and to convey results and conclusions as if the estimate captures *the* effect researchers are after.

An assumption underlying the meta-analysis is that the studies included provide information about a common underlying effect; in this case a multiplicative reduction in the odds or probability of child (under-5) mortality due to water treatment. For example, Kremer et al. (2021) describe results with statements such as "water treatment reduces the odds of all-cause child mortality by..." and "estimate the intention-to-treat (ITT) odds ratio impact of interventions to improve water quality on child mortality" and "...methods suggest that water treatment reduces the odds of all-cause child mortality by 25.9% ..." This language implies the quantity obtained from the meta-analysis represents a common underlying reduction in odds (and probability) of all-cause child mortality due to any water treatment; the assumptions needed to justify such statements are currently not explicitly addressed in the manuscript.

In the case of Kremer et al. (2021), the ten chlorination RCTs used in the analysis have substantial differences in their designs. As reported in the manuscript, "the age at which children were enrolled, as well as the periods for which they were followed, varies across studies, and some studies didn't collect data either on younger (<6 months) or older (> 2 years) children." The treatment and control conditions also differed. Having diversity among studies in terms of locations and time frames can help justify generalization of results to different regions and times if the collection of studies is large; however, for a small subset of studies, the heterogeneity actually creates challenges in terms of understanding and explaining what the combined estimate actually represents and whether it is worth such focus (as opposed to putting effort into explaining or modeling heterogeneity).

For Kremer et al. (2021), the small number of studies and design differences raise questions about what differences in what is being measured in each study, whether the estimates should be combined, and/or what a combined quantity might actually represent. In the context of a collecting data for a single study, suppose 10 pairs of children were chosen to participate in a study, with one in each pair randomly allocated to the treatment. Now, suppose the treatment and control conditions varied among the 10 pairs, as well as ages of the children, and how long they were followed. In the case of a single study, it would be common to criticize ignoring the differences in design among the ten pairs unless the decision to ignore was explicitly justified and explained. The underlying issue is the same for meta-analyses, yet researchers are rarely asked to justify

that the combined estimate is meaningful, even though this is still a key part of knowing how to interpret the results.

2.2.3 More on implications of heterogeneity among studies

The differences in effort (monitoring time) going into the counts of mortalities, as well as larger context (e.g. age of children included, underlying disease, properties of the water, outbreaks, etc.) are sources of heterogeneity that are not explicitly accounted for in the analysis. It is implicitly assumed that the mortality counts are data generated under the same design and as consistent measures the same underlying odds ratio (or relative risk). The within-study comparisons between treatment and control groups are meaningful in terms of *comparing* the odds of mortality between the two groups *within* the specific context of each individual study. However, to justify combining across studies the estimates should represent the same underlying effect, or an argument should be made that the combination of the different designs and contexts is what justifies the general statements. While there is no clear correct approach to this problem, the important part is acknowledging the potential limitations, opening discussion, and explicitly providing arguments and justification.

To start to address this problem, it can be helpful to first ask "odds of what?" for each study. For example, Crump (2005) provides an estimate of the reduction in the odds of mortality within 20 weeks after the start of their specific water treatment in the region of Kenya, with the particular aged children, and given other conditions and sources of mortality in the region. The study provides an estimate of the reduction in those odds associated with the water treatment, and each study differs in this regard. Explicitly discussing these differences opens the door to discussing the additional assumptions needed to get to the desired statements - such as the odds ratios associated with chlorination are constant over ages, constant over different implementations of the program, constant over different water properties, and constant across different regions with different underlying causes of childhood mortality.

Graphical exploratory data analysis could be used to plot the estimated odds ratios by design variables, such as summaries of the ages of children monitored, number of weeks monitored (see Figure 1), or variables capturing potentially important sources of heterogeneity among the studies. The fact there are only 10-13 studies makes it difficult to interpret the plots, but that in itself is also important information relative to justifying assumptions. Figure 1 provides an example investigating the estimated odds ratios relative to the number of weeks the children were followed. If the underlying odds ratio is same for all studies, we would expect to see a lot of variability among the estimated odds ratios associated with fewer weeks of monitoring and more stability in the estimates for the longer monitoring times (the reason for this is explained in more detail in the next section). As expected, we see the small variability among the three studies with over 100 weeks, but it is surprising how small the variability is for the four studies with the fewest weeks, as well as how relatively close to zero the estimated odds ratios are for those studies. Therefore, it would be worth discussing differences in designs or background situations for the four studies with the fewest weeks of monitoring and if there was anything in the study selection process that could have contributed to this.

An alternative to going after a single common effect may be to use statistical modeling to explicitly model how effects might differ for different regions, treatments, ages, etc. and then use that model to obtain context-specific predictions and associated statistical uncertainty. Such an approach would require additional data, but may be easier to justify and more useful in the long run if resources exist to support it. Unfortunately, it's common to start from the perspective that a single common effect is the goal, which isn't often challenged or discussed in reports on meta-analyses. In my experience, discussions of heterogeneity among studies often turn to a focus on fixed vs. random effect modeling, rather than stepping back to consider the meaning of an overall estimate relative to the differences among studies and expected heterogeneity in underlying effects. Random effects modeling is often described as a way to deal with heterogeneity, but the overall mean is still typically interpreted as *the* effect and such an interpretation should still require justification.

2.2.4 Possible implications of different lengths of monitoring

The different lengths of monitoring bring up additional issues regarding estimation of statistical uncertainty. Again, different aged children were monitored for time periods that varied across the studies (one study used

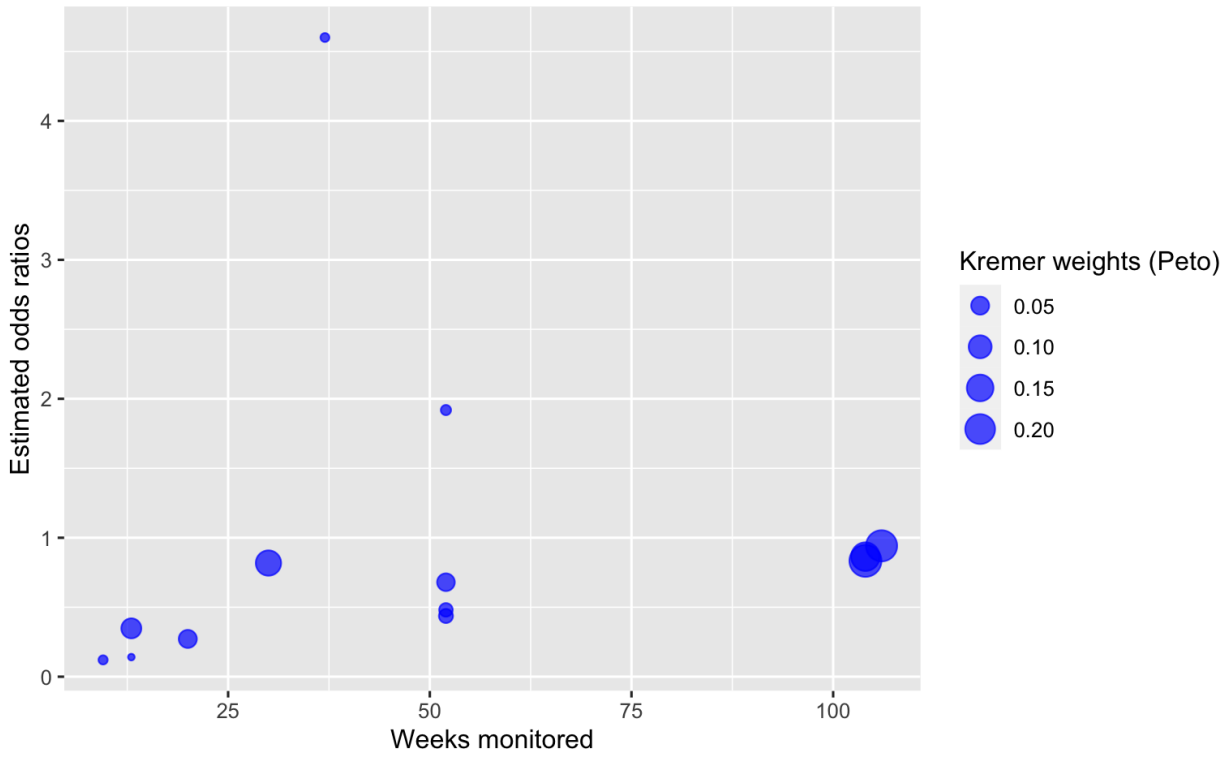

Figure 1: Exploratory plot providing an example of visualizing estimated odds ratios (Kremer et al. (2021) Peto ORs) against the number of weeks the children were monitored, with size of points representing the the weights used in one of the meta-analyses. Similar plots can be made for other variables that differ across studies.

9.5 weeks; two used 13 weeks; one each used 20, 30, and 37 weeks; four used 52 weeks; two used 104 weeks, and one continued for 106 weeks).

Odds ratios estimated from data collected over shorter time periods of monitoring are expected to exhibit more variability, for a given number of individuals and probabilities of mortality. For a study over relatively few weeks (e.g. 9.5), few deaths are expected in either the control group or the treatment group and the observed odds ratios are expected to jump around due to the small mortality counts. For example, with one death in the control group, going from no deaths in the treatment group to one death in the treatment group changes the observed odds ratio from 0 to 1, which is a substantial jump especially in terms of how it would be interpreted. This inherent variability in the odds ratios for small counts is important to consider and is not fully incorporated into the reported meta-analysis. Basic models for meta-analysis assume the effort going into each mortality count is the same, and while meta-analysis weights based on inverse variance will account for variability expected due to both the sample size and the size of the counts, it does not explicitly count for the differing efforts. This can manifest as giving studies based on few weeks more weight than they potentially deserve, as well as understating statistical uncertainty in the combined odds ratio. The Kremer et al. (2021) analysis gives largest weights to the three longest studies, as would be expected, and relatively small weights to two of the shortest studies. However, two of the four studies under 25 weeks had relatively large weights and small odds ratios. It would be worth looking into this in more detail and/or considering incorporating length of study into the weighting strategy, if it makes sense to use smaller weights for the studies that put less effort into the counts.

To explore the seriousness of the issue on reported statistical uncertainty (confidence/credible intervals), I simulated many random realizations of mortality counts from the binomial probability distribution, using the weeks and sample sizes consistent with the Kremer et al. data and assuming constant probabilities of mortality (and therefore constant odds ratio) over time and studies. The variability introduced into estimated combined odds ratios from the studies with shorter monitoring periods is substantial, raising concerns that the Kremer et al. (2021) confidence/credible intervals may be severely under-stating uncertainty in the estimates. At the very least, these aspects of the data and the assumptions necessary to get to the final statements should be discussed in the manuscript to help potential users of the results gauge the meaning and trustworthiness of the point estimates and reported intervals.

2.2.5 Possible implications of censoring relative to age 5

Choosing under-5 child mortality as the quantity of interest introduces another issue relative to how data are collected. The severity of the issue again depends on reasonableness of assumptions used to get around it. Often, a cross section of children of different ages are included in the study, with new births rolled into the study and children reaching 5 no longer monitored. For the study lengths included in this meta-analysis, the longest children were watched is a little over 2 years, therefore, children over 3 years old at the start would be followed until they reach, or could have reached, age 5 and all other children will not reach age 5 during the study. This makes sense logistically and can give reasonable estimates of under 5 mortality as long as all ages are monitored relative to their proportion in the population and assuming the outcomes for children who reached age 5 early and weren't subject to the intervention for as long were expected to have the same probabilities of death as if they would have been followed from an earlier age. If the intervention is expected to have an effect immediately and the effect is expected to stay constant over the length of the intervention, this could help with justification, though the same issue raised in the previous section can crop up again here if there are a lot of children who are almost 5 included in the study, or if the composition of ages of children in the control vs. treatment groups vary substantially.

Considering the individual children over time, rather than using a cross-sectional view, means that the data for children who are not tracked until they are five are technically censored (*i.e.*, the study ends before the outcome of interest is actually observed, so survival over the time period actually of interest is unknown for those individuals). Again, the authors can work to justify the assumption that the odds ratios would remain constant if the children were followed to the age of 5 (same number of additional children die in the treatment and control groups), and/or other reasons why the cross-sectional approach should closely approximate a longitudinal approach. There is no clear answer here, but it is worth addressing the complication and why individual level survival modeling is not necessary to obtain the desired estimates.

In short, the differences in ages and weeks monitored are important characteristics of the studies and data and deserve attention - whether through more sophisticated modeling to explicitly account for the differences, or addition of language to identify and justify assumptions underlying the current approach.

2.2.6 Generalizability to other regions and time periods

As with any analysis, the extent to which results should be generalized depends on what the data included can reasonably be assumed to represent in terms of new times, locations, and other contexts, as well as on what assumptions or arguments were used to arrive at the results. The Kremer et al. (2021) analysis uses a collection of studies that covers a wide geographical range and people with very different underlying living and health conditions. As previously stated, while heterogeneity can potentially help justify generalizations if the number of studies is sufficiently large, it can also introduce limitations when the number of studies is small, as is the case for Kremer et al. (2021).

The language in the manuscript implies broad applicability of the estimates as the reduction in probability of mortality for children due to water treatment in any region or time frame. For example, the authors demonstrate use of the results to calculate a "cost per DALY averted due to water treatment" without restrictions or cautions about potential limitations and use general statements such as "25% to 32% improvement in child survival." These types of statements are common when scientific results are translated to media reports, but having them in the scientific article to begin with can further contribute to problems of reporting over-stated and/or misleading descriptions of results relative to what is actually supported by the data and methods of the paper in question. Again, this concern can be at least partially mitigated through careful attention to wording and better justification for sweeping conclusions that do not directly follow from the data without additional assumptions.

Additionally, the authors should be very clear about when it is appropriate to use the general term "water treatment," as compared to the more specific term "chlorination." Unrestricted use of the term "water treatment" to describe results does not appear justified from the analysis.

These criticisms about generalizability and over-statement of results do not imply the results from the Kremer et al. (2021) analysis are not useful *conditional* on the studies included and any added assumptions, but the generalizations do not come for free simply because a meta-analysis was performed. Potential limitations and restrictions to the scope of inference should be articulated and discussed in depth before results are relied upon by others to represent something more than is justified. More attention could be given to nuances and limitations of the meta-analysis before moving on to recommend or demonstrate use of the results for subsequent calculations related to costs and lives saved. That is, while future applicability of the estimates is reasonable to discuss, it should come after more basic discussions of limitations and cautions.

2.2.7 Sensitivity analysis and uncertainty

The point estimates obtained from the meta-analyses appear reasonable *given* the studies they are based on and methods used to combine the information, assuming the meaning of the reported effect can be adequately explained and justified. Kremer et al. (2021) also reports on a fair number of sensitivity analyses (called 'robustness' in the manuscript), but it should be acknowledged that these are also conditional on the sets of studies originally included and excluded. The sensitivity analyses lend confidence that the results are not overly sensitive to decisions made after arriving at the 10-13 studies, but do not suggest a general "robustness" of the results beyond that; that is, they do not justify the extent of generalizations made throughout the paper.

Even if the point estimates are reasonable *given* the data included, the uncertainty captured in the reported confidence/credibility intervals is expected to be under-stated, and using point estimates while ignoring associating uncertainty quantified through the statistical analysis should be done with caution, even if just interpreted more qualitatively. The Kremer et al. (2021) manuscript could benefit from a meaningful discussion of sources and magnitudes of uncertainty within the research context and relative to how the authors are suggesting results can be immediately used to support other calculations related to the expected effects and costs of water treatment interventions.

2.2.8 Additional information and data

There is good reason to rely heavily on RCTs in terms of estimating potential causal effects of water treatment for a particular context; comparing a treatment to a control at the same time and in the same location is important to control for confounding variables and reduce variability that would be difficult to account for using observational studies and statistical modeling. However, this does not mean that observational studies or before-after studies at one location do not provide any information. An important use of statistical modeling can be to adjust for variables that can't be (or were not) controlled for in the design, though the success of this approach depends on data available to make the adjustments and requires more modeling effort. Therefore, I see results from observational studies on mortality as still providing valuable information that should be used as part of assessing the entire state of knowledge on the topic, particularly for those locations where the interventions will be carried out. Given the larger than expected magnitude of the estimated effect of chlorination from the Kremer et al. (2021) meta-analysis, as well as the substantial uncertainty, other sources of relevant information should be incorporated into the evaluation process, even if not combined quantitatively.

2.2.9 Publication bias testing and correction

The authors employ statistical tests for publication bias and find that "neither Egger's, Begg's, or Andrews and Kasy's tests provide evidence of publication bias" and report later that "Possible publication bias was examined with inspection of funnel plots and the use of Egger's test." While these methods can be useful in helping to assess the extent of publication bias in meta-analyses, they are limited in the usefulness particularly when applied to a small number of studies, such as in Kremer et al. (2021). Recommendations found in Cochrane guidance do not recommend use of the methods with less than 11 studies, as there is very little information to assess assymetry in the funnel plot. A large p-value from such a test does not imply evidence of no publication bias; that is, the results may be consistent with what would be expected under no publication bias, *as well as* what would be expected under some extent publication bias. It would be difficult, if not impossible, to argue that there is no publication bias in this context. Therefore, in this case, I find the decision by GiveWell to assume the presence of publication bias and use a bias-corrected estimate to be a reasonable way to proceed, particularly given the large uncertainty and unexpectedly large magnitude of the unadjusted point estimate.

3 Conclusions

3.0.1 Potential for future work

As mentioned previously, it may be advantageous to attempt to explicitly model some of the sources of heterogeneity among studies, with the goal of making predictions for specific scenarios where interventions may be deployed, rather than developing a single relative risk to use in any situation. Additionally, more sophisticated Bayesian modeling could help address the challenges of combining many sources of information, making it possible to formally include prior information from experts, as well as many other studies beyond the RCTs included in the Kremer et al. (2021) meta-analysis. However, it would be a big undertaking and may not be worth the time, particularly if only point estimates are going to be used to inform decision making.

Regardless of whether more sophisticated modeling approaches are ultimately employed, I believe the work could benefit from collaboration with a statistician with expertise in logistic regression, survival analysis, meta-analysis, and Bayesian modeling. While the analyses can be easily carried out using a computer and existing code, improvements to the modeling, associated descriptions, and justification of assumptions could greatly improve the credibility, trustworthiness, and ultimate usefulness of the work.

3.0.2 Recommendations

Given the limitations and sources of uncertainty discussed in this review, the results from the meta-analysis described in Kremer et al. (2021) should *not* be taken as the primary or most valuable source of information

simply because they come from a meta-analysis of RCTs. Instead, the results can be incorporated as additional information describing the potential effects of chlorination on mortality of children of different ages based on *these* 10 studies. It is not clear, without further justification, how the results should be generalized to other situations, and it is not clear that the estimate from the meta-analysis should be given substantially more weight than other sources of data and expertise.

As already described in the GiveWell report, the results should be reality checked compared to other empirical knowledge and theoretical expertise regarding potential effects of chlorination on child mortality in low and middle income countries. While it is attractive to base cost effectiveness analyses on a single number coming from a trustworthy quantitative analysis, that calculation carries with it the limitations of the analysis creating it.

The additional information and adjustments incorporated by GiveWell in their report appear adequately justified, with clear descriptions of the decisions and rationale behind them. Regardless of method, it seems that uncertainty in potential effects will likely remain greater than desired to support investment decisions based on a single number; therefore, I recommend providing a range to potential donors rather than relying on a single number that implies more certainty and trust than is justified. I also suggest reducing the precision used in reporting the results of the meta-analysis to more honestly reflect the inherent uncertainty (e.g., "about 25 %" is more appropriate for the context than reporting 25.9 %).