Hi Keith

I have written responses to your “critical comments” post, and then to the “what am I missing” discussion. What you wrote is in italics. I have not commented on your review of Rubin’s conceptual ideas. The responses about the goals of the paper reflect earlier discussions with Julian Higgins and Thomas Lumley. The other comments reflect my thoughts alone.

I urge you to re-read the paper, but bearing in mind the goals for it I have stated below, frequently. The paper may not address all your concerns with the practice of meta-analysis – it doesn’t address all of ours – but I hope it reduces your level of frustration with it.

Critical Comments:

Ken: Unlike some of your other papers this one I find unconvincing if not even wrong headed.

I don’t think your comments below bear out the “wrong-headed” epithet here. But perhaps I am missing something.

Yes “answers a question that is relevant to the scientific situation at hand” is the goal but Peto and Peto like automatic/thougtless post-stratification of (reified) fixed sub-population effects are very unlikely to that credibly.

First of all, this is ungrammatical, it does not make sense and I am not able to guess what you mean. Pulling what I can from it, it seems confused: I strongly doubt either Peto would be in favor of “thoughtless” application of any method – and neither am I, nor do I think the JRSSA paper encourages this. So this is a mystery to me. I’m sure this is frustrating for you but it’s also frustrating to see our work labelled “wrong-headed” without comprehensible reasoning as to why.

I believe a better route forward, if there is knowledge of what to post-stratify on to get a stable over time and place population of interest would be MRP with an informative prior on effect variation. If not just a random effect model with an informative prior on effect variation.

The initial stress on a “better route forward” here seems to miss the main point of our JRSSA paper. Fixed-effects meta-analysis answers a question, and one that can be asked legitimately regardless of assumptions of homogeneity of the underlying effects. (Note that whether it is an **optimal** question to ask is context-specific, and it’s certainly not the only question one could ask in most applied settings). There are many mis-statements about fixed-effects methods in the meta-analysis literature; our goal was to point them out and suggest how to instead think about which question(s) were **actually** relevant. The only “route” I see laid out in our paper is to have users think about fixed-effects methods, without being misled by those mis-statements, and choose for themselves whether it is relevant to their work. Fixed-effects is not the only tool, nor the only one we want people to use.

I am not sure what you mean by a “stable over time and place population of interest”, in the paper we have tried to use “population” with its technical meaning – in which case stability applies to the estimates produced using data from a population, not to the population itself. And estimation of the fixed effects parameter is perfectly possible under a prior that assumes exchangeability (i.e. motivates standard random-effects analyses), although we didn’t get into

this in our paper. So I'm not sure that there is quite such a contrast between what you might prefer and the various arguments presented in our paper.

It is the avoidance of informative priors that drives the desperate holy grail quest to make sense of varying effects as fixed – see Dan's new post here http://andrewgelman.com/2017/09/05/never-total-eclipse-prior/

I don't recognize anything from the paper as "desperate" or quixotic – nor do you say how these criticisms apply to what we wrote. This is again frustrating. Do however note that the fixed effects (plural) assumption just says "look, we've got a lot of different studies here, each estimating an underlying parameter". The effects can be different but they are not viewed as varying randomly, i.e. being sampled from some population.

The idea that a trial had a certain subset of the population and that the effect is fixed is that – is just a hopeful untested superstition supported by not have access to that variation – surely it will vary by time and place (though perhaps not importantly). Perhaps, inbreed lab animals in the same lab with same technician – but there we do see above SE variation. Trial are attempts to things correctly as possible, there is always some slippage. The Rubin paper you referenced was very clear about methodological in addition to biological variations.

I have trouble, again, guessing what you mean here; a "hopeful untested superstition supported by not have access to that variation" makes no sense. Perhaps you also mean "supposition"? From what I can glean, the definition of "population" appears to be causing confusion. I interpret studies as drawing data from an underlying population, in which the parameter value is fixed. If parameter values vary this means one is considering different populations, e.g. median height in men vs women. If you don't care for this notion, fine, but it's assumed in everything we do in the paper, and in overwhelming amounts of the meta-analysis and broader statistical methodology literatures. I hope you understand the need to write papers using frameworks that are familiar to most of the audience.

Was not aware that this fixed versus random effects was such a big deal in genetics.

It has been an issue in high-throughput analyses. Fixed-effects is becoming the standard approach. The work by Danyu Lin and co-authors (extending earlier results by Olkin and others) on the connections between fixed-effects approaches and combined-data analyses have been pivotal in this; in genetics if one could pool the data one would and the analyses fitted to combined data are very close to those in Danyu's results. Backing this up, recently we have started to pool large genetic datasets, and such analyses are being done, routinely.

Though even with Mendelian randomization I do think you mostly "speak of no non-random error".

I don't speak of this, nor do I know who you are quoting.

Or maybe I missed something.

Quite possibly.

What am I missing and what will this paper likely lead researchers to think and do?

Posted by Keith O'Rourke on 5 October 2017, 3:00 pm

In a previous post Ken Rice brought our attention to a recent paper he had published with Julian Higgins and Thomas Lumley (RHL). After I obtained access and read the paper, I made some critical comments regarding RHL which ended with "Or maybe I missed something."

This post will try to discern what I might have missed by my recasting some of the arguments I discerned as being given in the paper. I do still think, "It is the avoidance of informative priors [for effect variation] that drives the desperate holy grail quest to make sense of varying effects as fixed". However, given for argument's sake that one must for some vague reason avoid informative priors for effect variation at all cost, I will try to discern if RHL's paper outlined a scientifically profitable approach.

I don't recognize the desperation you suggest here, nor do I think reasons for not using informative priors need be "vague". As you'll have noted in the paper, we make "standard" assumptions, including that large-sample approximations are near-enough accurate not to be a big concern. This doesn't apply to every situation but is also not uncommon. A typical consequence of that assumption is that priors would have to be very informative to affect the output much, with much more precision coming from the prior than the data. What does matter is what we choose to estimate; should it be the fixed-effects parameter? And/or something else?

However, I should point out their implied priors seem to be a point prior of zero for there being any effect variation due to varying study quality…

This is a misconception. The paper goes to considerable lengths to allow for underlying effects to differ. See Figure 2b and related text.

…and a point prior of one that the default fixed effect estimate can be reasonably generalized to a population of real scientific interest.

This is also a misconception. The parameter estimated by fixed-effects analysis can indeed by interpreted in terms of a population, as we describe. Whether this population is of scientific interest is context specific; see Section 4 for explicit discussion.

In addition to this, as I think the statistical discipline needs to take more responsibility for the habits of inference they instil in others I am very concerned what various research groups most likely will think and do given an accurate reading of RHL?

As mentioned above, we hope the paper provides an accurate understanding of fixed-effects estimates and the inference that goes with them. This is a "habit of inference" for which I am happy to take responsibility.

Succinctly (as its a long post) what I mostly don't like about RHL's paper is that they seem to suggest their specific weighted averaging to a population estimand – which annihilates the between study variation – will be of scientific relevance and from which one can sensibly generalize to a target population of interest.

This is a misconception. The weighted average **may** be of interest, or may not, and we explicitly say this. What you "mostly don't like" (perhaps you do like it sometimes?) is something we did not write. Furthermore your comments about "annihilating" between study variation also does not reflect what we wrote. By reporting a weighted average of effects we make no pretense that study variation does not exist; see Section 5.

Furthermore it is suggested as being widely applicable and often only involves the use of default inverse variance weights. Appropriate situations will exist but I think they will be very rare.

While I suppose reasonable people can differ on this, in our experience they are not rare.

Perhaps most importantly, I believe RHL need to be set out how this will be credibly assessed to be the case in application. RHL does mention limitations, but I believe these are of a rather vague sort of don't use these methods when they are not appropriate.

As noted in the paper (section 6) we note that users could consider re-weighting to other populations, and give an example. I agree with you that other tools could be provided here. But doing so was not a goal of the paper – recall we just want readers to understand fixed-effects analysis better – so this criticism is like complaining that the bicycle pump we have provided does a terrible job of peeling your potatoes.

That is seemingly little or no advice for when (or how to check) if one should use the publication interesting narrow intervals or the publication uninteresting wide intervals.

I don't view it as an either/or question. Use fixed-effects analysis if it answers a relevant question, use random-effects approaches if they answer a relevant question. If neither answers your question, use neither. If you don't know the question, don't expect the information in the meta-analysis to tell you.

First, I will review work by Don Rubin that I came across when I was trying to figure out how to deal with the varying quality of RCTs that we were trying to meta-analyse in the 1980,s. It helps clarify what meta-analyses should ideally be aiming at. He conceptualized meta-analysis as building and extrapolating response surfaces in an attempt to estimate "true effects." These true effects were defined as the effects that would be obtained in perfect hypothetical studies. I referred to this work in my very first talk on meta-analysis and RHL also referred to this paper on it – Meta-analysis: Literature synthesis or effect-size surface estimation? DB Rubin – Journal of Educational Statistics, 1992, though in a very limited way. I think I prefer Rubin's earlier paper "A New Perspective" in this book. I will then apply this perspective of what meta-analyses should ideally be aiming at to critically assess where RHL's proposed approach would be most promising.

Now, Rubin was building and extrapolating response surfaces out of a concern that "we really do not care scientifically about summarizing this finite population {of published studies we have become aware of) but rather "the underlying process that is generating these outcomes that we happen to see – that we, as fallible researchers, are trying to glimpse through the opaque window of imperfect empirical studies". He argues that to better do this we should model two kinds of factors – scientific factors (I often refer to this as biological variation) and scientifically uninteresting design factors (I often refer to this as study quality variation). Furthermore as we want to get at the underlying process, we need to extrapolate to the highest quality studies as these more directly truly reflect the underlying process.

Using the notation X for biological factors and Z for study quality factors he is "proposing, answers are conditional on those factors that describe the science [biology], X, and an ideal [quality] study Z = Z0. That is where we are making inferences." If there is a lot of extrapolation uncertainty – then that's the answer. Not much to learn from these studies, they are not the ones you are looking for, so move on.

Now his work has been primarily conceptual as far as I am aware. Sander Greenland and I tried to see how far we could take modelling and extrapolation in a paper entitled "On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions." Unfortunately no one seemed willing to share a suitable data set for us to actually try it out. Ideally, such a data set would have a reasonable number a studies that have been adequately assessed for their quality (which we defined as whatever leads to more valid results which likely is of fairly high dimension with the quality dimensions being highly application-specific and hard to measure from published information). Given such requirements are quite demanding, I don't think they will be met in the published clinical research with primarily only access to the publications. Perhaps though in a co-operating group of researchers studying a common disease and treatment prospectively conducting studies where they should have access to all the study protocols, revisions and raw data. Or perhaps a prospective inter-laboratory reliability study. In summary, there is always a need to adequately model a X, Z surface and extrapolate to Z0, with this being ignorable only when all studies are close enough to Z0.

So we are moving on to just considering RHL's method in consistently high quality studies – they are not the methods to use when there is varying quality.

First, fixed-effects meta-analysis is not "RHL's method", we did not devise it. Issues of quality pertain to issues, discussed above, of whether the overall population is of interest. It does not seem helpful to repeat them again.

A rare situation but it can happen.

Not, again, in our experience.

I believe it would exclude most meta-analysis done within for instance the Cochrane Collaboration. Not sure if RHL would agree. Even within this restricted context of uniformly high quality studies, I think it will be helpful to give my sense of the science or reality of multiple high quality studies as even that can be tricky.

Each study likely will recruit an idiosyncratic sample of patients from the general population – it will be highly selective (non-random) and not that well characterized. This can be seen in the variation of patient characteristics and for instance the varying control group outcomes in the trials. There will be a restrictions on eligibility criteria. However, within that there there can still be much variation. So each study will have a idiosyncratic sample of patients recruited from the general population which provides a relative percentage in that study of the total studied population. RA Fisher used the term indefinite to refer to this type of situation where we have no means of reproducing such differing sub-populations at will. The same investigators at a later time or in a different city would unlikely be able to recruit the same mix of patients (i.e. the recruitment into clinical trails is pretty haphazard in my experience). Because of this I don't how the sub-population could even be adequately described nor the relative percentage of this sub-populations in a population of interest ever be determined.

Now, the relative percentage of the total studied population of these study sub-populations will likely differ from the percentage of these same idiosyncratic sub-populations in the general population or some targeted population of interest. Sample sizes in conducted trials is largely determined by funds available, recruitment abilities of the trial staff, other trials competing to enrol subsets of targeted patients at the time, etc. Because of this I don't see why it would be expected that the relative percentages of the various study recruited patients (of the total studied population) would approximately equal the relative percentages of the various patients in a general or target interest population.

Please see earlier responses on "populations". I agree that the exact mix of patients will not be known, but "Table 1" of almost all study reports will give a good indication of the study participants – this is its whole purpose. Working with skilled epidemiologists (as I'm lucky enough to do) it has been my experience that they are **very** alert to population differences, and the likely impact on results.

Now in a very convenient case where the only patient characteristic that drives treatment effect variation is say gender – we will know the relative percentages in the study population and likely any targeted population (with negligible error) and post-stratifying (weighting to match to a target population gender proportions) will likely be straightforward. RHL provides an appendix which actually carries this out for such an example. But in realistic study settings, I have no idea how the needed weights could be obtained. That is, we will usually just have idiosyncratic samples of patients in studies without much if any knowledge of their make up or their relative percentages in targeted populations or what drives the biological variation (given there is variation).

The choice of sex for post-stratification was chosen for conceptual simplicity. I agree that more detailed guidance for use of multiple variables in this way could be given, but as it is quite far from the main argument of our paper we chose not to do so. I recommend Thomas Lumley's book on Complex Surveys for general discussion of reweighting strategies.

Now it is clear that we do have the study population and so it should be fairly direct to assess the average treatment effect in the studied population. Here between study scientific variation (i.e. study population variation) can be taken as fixed and ignored. However, I would argue this question is not one of scientific relevance (that RHL is primarily interested in) but rather a practical economy of research question – is it worth continued study of this intervention to get a better sense of what that effect would be in a targeted population and what studies should we do to get a better sense. This perspective goes way back to Fisher with his careful discernment of what variation should and should not be ignored for questions of scientific relevance.

It is not clear what "this question" is here. As mentioned above, our paper does not condone ignoring heterogeneity, though heterogeneity may not be described in every form of inference we use in a meta-analysis setting.

As an aside, I have thought a fair amount about these issues as I presented a similar weighting argument to get an uninteresting average magnitude effect estimate and an interesting average sign effect estimate – the latter just being positive or negative. I presented the argument to a SAMSI Meta-analysis working group in 2008. I recall Ken and Julian being there but I am not really sure and they likely do not remember my presentation either.

I was there but do not recall it.

Jim Berger criticized the population that was defined by the inverse variance weighting as being non-existent or imaginary. Now I had thought the assumption of the treatment effect being monotonic would make that criticism moot, but I was not sure at the time. Richard Peto often insisted he was justified in making such an assumption and I was taking that as a clue. If the effect is only positive or negative, then an average effect in any population real or counterfactual – no matter how uninteresting – would enable the sign to be pinned down for any other population. I later discussed this with David Cox via email and he argued that monotonicity was a very questionable assumption. Furthermore, if I actually wanted to make such an assumption, why not assume a treatment variation distribution on the positive line? So I abandoned the idea.

This is very confusing. You have not said what any treatment effect is monotonic with respect to, nor can I guess.

Now some specific excerpts from RHL that I may benefit from some clarification:

RHL > "we discuss in detail what the inverse-variance-weighted average represents, and how it should be interpreted under these different models" and "[the] summary that is produced should be interpretable, relevant to the scientific question at hand and statistically well calibrated … controversial issue that we aim to clarify is whether ˆβ in equation (1) [inverse-variance-weighted average] estimates a parameter of scientific relevance."

I definitely like these promises but I don't see them being explicitly or adequately met in the paper.

Not meeting this is unfortunate, but please see above for some issues that may not have been understood correctly.

RHL > "we restrict ourselves to the situation of a collection of studies with similar aims and designs, free of important flaws in their implementation or analysis. (See Section 6 for further discussion.)"and then in Section 6 "when studies do not provide valid analyses, either because of limitations in the design and conduct of the study, or because, after data collection, post hoc changes are made to the analysis, but reported analyses do not take these steps into account… If in practice these procedures cannot be avoided, accounting or the biases that they induce is known to be difficult…"

The challenge here is that its not clear what is meant by important flaws and when they are present they seem to be suggesting not much can be done. For instance, what percentage of the meta-analyses in the Cochrane collection of meta-analyses would have such flaws – 10% or 90%? Would one or more entry in the risk of bias tool help sort this out?

No, we do not address these challenges – but nor was this the goal of our paper. Please see the comment above on bike pumps and potato peeling.

RHL>"[inverse-variance-weighted average] estimates a population parameter, for the population formed by amalgamating the study populations at hand. … in the overall population that amalgamates all k individual study populations, define ηi as the proportion of the population that comes from study population i."

I am not sure whether RHL mean to define ηi as the proportion of the total studied population that comes from study i or the proportion of a general or population of interest that matches study i's sub-population. The following two claims suggest it is the first.

Yes, the former.

RHL>"we see that ni/Σni [the proportion of total study population in study i] is consistent for ηi" and "proportions ηi are known with negligible error"

The the population of scientific relevance is the second and so how to we get to that – just assume the proportion of the total studied population roughly equals a population of interest? That surely needs some justification?

Given the didactic goal of the paper, the issue is not so much whether such a population is of interest, but just the realization that the analysis is informing us about such a population.

RHL>"It remains to discuss the scientific relevance of β; the use of this specific weighted average is described … general results given in Section 3.3."

I very much agree that it does remain and RHL claim it will be in Section 3.3.

But Section 3.3 is just an inverse-variance-weighted average view or recasting of general regression into sub-pieces – which I can not see as addressing scientific relevance. As if general regression was the definition of scientific relevance!?

See above for my comments on Danyu Lin's work and its impact in genetics. The point of Section 3.3 is to show the very close connection between fixed-effects meta-analysis and what most investigators would do (and do actually do) if/when it's possible to pool all the data and do a single analysis. It's scientifically relevant because it's what very many scientists would do if they could pool data.

Rather, I would (and have) argue(d) its just an analogue of various ways to factorize the joint likelihood of studies with various choices of (in RHL terms) identical versus independent parameter assumptions.

"Analogue" here is such a vague term that I don't think it's helpful. Moreover, joint likelihoods did not seem a fruitful way to convince a general audience of our main arguments, which is why we did not present them.

*For example in the simplest example of regression of a single x through the origin, specifying an identical slope parameter for all studies but independent within study variance parameters different for each study (that is each study gets its own variance parameter). The usual regression involving all the study's raw data in one regression being expressed as Normal(beta,sd1, sd2,…,sdn,x,y) that is being rewritten as Normal(beta,sd1,x1,y1) * Normal(beta,sd2,x2,y2) * …. * Normal(beta,sdn,xn,yn). As these individual study likelihoods are exactly quadratic (given RHL's assumptions) they can be replaced with inverse variance weighted individual study beta estimates. So what?*

Given that we didn't use likelihoods I don't see the relevance of this question.

RHL>"the fixed effects meta-analysis estimates a well-defined population parameter, of general relevance in realistic settings. Consequently, assessing the appropriateness of fixed effects analyses by checking homogeneity is without foundation— … Both in theory and in practice, the argument is not tenable and should not be made."

I think this a very strong claim given what is and isn't in the paper.

Please review Equations (1) and (2) in the paper to see how homogeneity is not needed for the fixed-effects meta-analysis. This is quite contrary to what's recommended in much of the literature, which is why we are so blunt. Also note that (as per Section 5) we do not rule out considering heterogeneity alongside estimates of parameters describing averages.

Furthermore – checking homogeneity is just checking some of the assumptions of the data generating model. If one is not making a common parameter assumption but rather a independent parameter assumption, does that make model checking impossible?

Not impossible but not needed, for inference on the average parameter we have described.

That would be convenient. The minimalist of data generating model assumptions for meta-analysis is there not being apples and oranges. Something is being taken as common in these multiple studies – is that in error? How would one ever know?

Again, you seem to be overlooking Section 5, titled "Dealing with heterogeneity alongside a fixed-effects meta-analysis".

"For example, if the subjects in the studies contributing to the meta-analysis are representative of an overall population of interest, the fixed effects estimate is directly relevant to that overall population … If, however, the sample sizes across studies vary so greatly that the combined population is unrepresentative of any plausible overall population, then the fixed effects parameter will not be as useful."

Maybe this explains what I am missing – RHL is only suggested the method be used when you know the mix of idiosyncratic study populations are actually is representative of an overall population of interest. That is if and only if the proportion of total study population in study i is consistent for ηi. How would one know that? How would one check that?

Rather than checking this assumption, I think it's more relevant to realize that **that is the population we are learning about**, in general, from fixed effects meta-analysis. If, based on contextual knowledge, you have a different overall population in mind, one would "check" by comparing the two. Given the strong context-specific knowledge needed this is not a major focus of our paper.

But RHL also claimed it was of general relevance in realistic settings so they are assuming in most realistic settings the proportion of total study population in study i is consistent for ηi? Or so it seems given this statement.

RHL>"Fixed effects meta-analysis can and often should be used in situations where effects differ between studies"

The “often” part here seems to be the problem. As noted above, it is (in our experience) often the case that the population described is of relevance to underlying scientific questions. We did not claim blanket “general relevance” as you imply.

Now for somethings I do fully agree with.

RHL> “However, if the random-effects assumption is motivated through exchangeability alone”

Yup – it is surely one of those very untrue models (a really really wrong model as the Spice Girls would put it) but which is useful in bringing in some of the real uncertainty – though admittedly seldom the right amount. That is why I once was claimed it was not the least wrong model.

I am glad we agree on something.

RHL> “Measures of heterogeneity should not be used to determine whether fixed effects analysis is appropriate, but users should instead make this decision by deciding whether fixed effects analysis—or some variant of it—answers a question that is relevant to the scientific situation at hand.”

I fully agree – but again its the banning of informative priors altogether – forcing there to be a discrete decision to either completely ignore or fully incorporate very noisy between study variation. Nothing in between! And with seemingly little or no advice to be had for when (or how to check) if one should use the publication interesting narrow intervals or the publication uninteresting wide intervals. This is the real problem.

What is the banning of informative priors? I see no connection between what you wrote and our statement on statistical measures of heterogeneity.