Published in final edited form as:

Epidemiology. 2015 September ; 26(5): 710–713. doi:10.1097/EDE.0000000000000359.

On Effect Measures, Heterogeneity, and the Laws of Nature

Orestis A. Panagiotou^a and Thomas A. Trikalinos^b

^aDivision of Cancer Epidemiology & Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD

^bCenter for Evidence-based Medicine, and Department of Health Services, Policy & Practice, School of Public Health, Brown University, Providence, RI

Doing a meta-analysis is easy.

Doing it well – well...

Ingram Olkin, at a leisurely breakfast on a sunny Sunday of this century

When it comes to the effectiveness and safety of treatments, does nature prefer multiplicative or additive effects—or something else?

Empirical surveys of meta-analyses comparing treatments for dichotomous outcomes suggest that the choice of effect measure is associated with how (in)consistent, or heterogeneous, study effects appear to be between them.^{1–3} A typical observation is that statistical tests for between-study heterogeneity are more often significant with a risk difference measure compared with a multiplicative measure, such as the odds ratio.^{1,2}

In this issue of *Epidemiology*, Poole et al.⁴ submit that we should not rush to interpret the aforementioned observations as indications that nature works on a multiplicative rather than additive scale, or as an implication of “a certain ontological property about the universe: about human persons and the exposures, treatments and outcomes they experience.” Much as the various measures of effect refer to different scales (e.g., the risk difference refers to the probability scale whereas the log-odds ratio refers to the logit-probability scale), so do all commonly used measures of effect heterogeneity. The scale-dependence of the analyses precludes one from readily comparing statistics (including heterogeneity *P* values) across different effect measures. All commonly used measures of heterogeneity are in this respect unfulfilling, and a scale-independent assessment of heterogeneity is needed.

We agree with Poole et al. that statistics do not readily transfer across scales, and outline a modest proposal for recasting models meta-analytical in a framework that might allow comparisons of analyses conducted in different scales. We second the authors’ call for developing novel and more meaningful measures and explorations of heterogeneity, even for

Correspondence: Orestis A. Panagiotou, Division of Cancer Epidemiology & Genetics, Room 7E602, 9609 Medical Center Dr., Bethesda, MD 20982. orestis.panagiotou@nih.gov; or Thomas A. Trikalinos, Center for Evidence-based Medicine, School of Public Health, Brown University, 121 South Main Street, 8th Floor, Providence, RI 02903. thomas_trikalinos@brown.edu.

The authors report no other conflicts of interest.

analyses that use a specific effect measure. We are similarly skeptical of bold interpretations of meta-epidemiological observations, and offer some higher lever critiques.

META-ANALYSIS IN THE GENERALIZED LINEAR MIXED MODELING FRAMEWORK

We recast meta-analysis in the generalized linear (mixed) modeling (GL[M]M) framework, in an attempt to reframe choice of effect measures and modeling of heterogeneity as tasks that are amenable to existing theory and tools. While, strictly speaking, this analytical framing is different from the analyses conducted in the meta-epidemiological studies mentioned above, it is for most practical purposes equivalent, and perhaps preferable.

Consider a collection of studies, indexed by $k = 1, \dots, K$, that compare an intervention versus a control with respect to a binary outcome. (For simplicity, we focus here on two-arm trials, but this discussion generalizes to multi-arm trials.) Provided that the definitions of the populations, the versions of the compared interventions, and the outcome are similar-enough in all K studies, to some of us it makes sense to learn the intervention effect across them. But first, let us define effect measures in a single study, and fix some notation.

Measuring the Effect in a Single Study

For the k -th study, write x_{k0} (x_{k1}) for the number of events and N_{k0} (N_{k1}) for the total number of participants in the control (intervention) arms. Write also π_{k0} (π_{k1}) for the respective population probabilities of events. We measure the effect of the intervention, δ_k as the difference between functions of the probability of events in the two groups

$$\delta_k = g(\pi_{k1}) - g(\pi_{k0}), \quad (1)$$

where $g(\cdot)$ is a function that is monotonic over the unit interval $(0,1)$. The common effect measures of interest are obtained for a suitable $g(\cdot)$. One obtains the risk difference when $g(\cdot)$ is the identity function, the log risk ratio when $g(\cdot) = \log(\cdot)$, and the log odds ratio when $g(\cdot) = \text{logit}(\cdot)$.⁵ Although the risk difference, risk ratio and odds ratio are the measures epidemiologists are most familiar with, an infinite set of measures can be defined, because infinite $g(\cdot)$'s exist. For example, epidemiologists who were experienced before we (Orestis and Tom) could walk might recall measures that are now long out-of-vogue, such as the “difference in the angles,” which is obtained by setting $g(\cdot) = \arcsin(2\sqrt{\cdot})$, the first-order variance-stabilizing transformation for binomial proportions, a measure that was popular before computational solutions for fitting generalized linear models (GLMs) became widely available. Engineers might be partial to differences in entropy, which can also be obtained with a suitable $g(\cdot)$. In general, people in different fields choose different effect measures (equivalently, different $g(\cdot)$'s) based on interpretability, mathematical and computational considerations, and perhaps, we shall argue, on model parsimony and data fit.

Using the notation in (1), we analyze study k by positing an observational model for the counts

$$x_{kt} \sim \text{Binomial}(\pi_{kt}, N_{kt}), \quad (2)$$

and a structural model for the parameters

$$g(\pi_{kt}) = a_k + \delta_k \cdot I(t=1), \quad (3)$$

where t indexes treatments, $I(t=1)$ is an indicator function returning 0 when $t=0$ (control) and 1 when $t=1$ (intervention), and a_k is the $g(\cdot)$ -transformed proportion of events in the control arm. For a single study, the GLM in (3) is saturated and fits data perfectly irrespective of the choice of the link function $g(\cdot)$. Thus, analysts would choose the effect measure (link function) that is easier to interpret, results in more power to detect a difference, or, more mundanely, is easier to fit a model with.

Measuring the Effect in a Meta-analysis

In a meta-analysis, the goal is to learn the effect measure across studies, and so, in addition to the within-study Equations (2) and (3), we must impose some structure *across studies*. This structure encodes a belief about how the different δ_k relate between them, and enables us to learn a common effect. A popular option is an equal effect (also termed “fixed” effect) model, which assumes that all studies have the same common effect :

$$\delta_k = \Delta \text{ for all } k=1, \dots, K. \quad (4)$$

Another popular option is to assume that the δ_k are exchangeable random effects, that is, are distinct but follow a (here, normal) distribution

$$\delta_k \sim N(\Delta, \tau^2) \text{ for all } k=1, \dots, K. \quad (5)$$

where Δ is the mean and τ^2 the variance (heterogeneity) between-studies.

Thus, equal effect meta-analysis can be written as a GLM with the observational model as in (2), and the structural model as

$$g(\pi_{kt}) = a_k + \Delta \cdot I(t=1), \quad (6)$$

where the a_k are nuisance parameters. Analogously, random effects meta-analysis can be seen in the generalized linear mixed modeling (GLMM) framework, with the observational model as in (2), and the structural model as

$$g(\pi_{kt}) = \underbrace{a_k + \Delta \cdot I(t=1)}_{\text{fixed part}} + \underbrace{\zeta_k \cdot I(t=1)}_{\text{random part}} \quad (7)$$

where $\zeta_k = (\delta_k - \Delta) \sim N(0, \tau^2)$. Note that the equal effect model in (6) is nested in (is the fixed part of) the random effects model in (7). Furthermore, neither model is saturated and it is possible for the random effects model to fit data better than the fixed effects one.

Choosing Between Effect Measures and Between Equal (Fixed) Effect and Random Effects Models

How can one check which effect measure (which link) fits data better? By recasting meta-analysis in the GLM/GLMM framework, we recast the choice of effect measures as an exploration of the “goodness-of-link” of alternative link functions $g(\cdot)$.⁶ Also, for a given link function/effect measure, we recast the choice between equal (fixed) effect and random effects models as an exploration of the “goodness-of-fit” between regressions that include versus not include the random effects part in (7) in the right hand side. Analysts are less familiar with explorations of the goodness of link functions, perhaps because in most applications one can improve model fit by changing the linear predictor alone (e.g., adding predictors, interactions, or nonlinear terms) without “messing” with the link function (see chapter 11.3 in McCullagh and Nelder⁷).

Starting with the most familiar, for a given link function, one can compare whether the random effects model in (2), (7) fits the data significantly better than the equal effect model in (2), (6). For example, if the models are estimated with maximum likelihood, one can use a likelihood ratio test to compare the two models that have the same fixed part, and differ by a variance component (the between-study heterogeneity). A complication stems from the fact that we test a nonnegative (variance) parameter on its boundary (on 0), but theory for such comparisons exists.^{8,9}

To illustrate an approach to assessing the fit of models with alternative link functions to data, consider Pregibon’s “goodness-of-link” tests.⁶ The idea is to construct parametric link function families, which degenerate to specific link functions for different values of their parameters.⁶ For example, the single-parameter link function family

$$g(\pi|\lambda) = (\pi^\lambda - 1)/\lambda \quad (8)$$

yields the identity link (and a risk difference measure) when $\lambda = 1$ and the log link (and a log-risk ratio measure) in the limit as $\lambda \rightarrow 0$. A goodness-of-link test can then be constructed as a score test under the null hypothesis of $\lambda = \lambda_0$, where λ_0 is a fixed value.^{7,10} Other link function families can be constructed, including step functions that choose between two (or more) specific link functions, instead of using a continuous parameter, such as the example above. Application of methods for evaluating the fit of the link function have been explicated for equal effect and for random effects models in the GLMM framework.¹¹

The descriptions above pertained to comparisons across models that differ only in the link function used, or in whether they include random effects or not. Comparing the fit of models to data across different link functions and across equal effect versus random effects models is more complex. Because the models are not nested, likelihood ratio tests cannot be used.¹² We do not know whether tests exist to assess the statistical significance of such comparisons. An option might be to resort to heuristic model comparisons conditional on data, using various information criteria (such as the Akaike and the Bayesian information criteria), or, in the Bayesian setting, through posterior predictive checks.^{13–15}

Once a preferred link, and hence the effect measure, that best describes the data in-hand has been identified, one can present the evidence in scales that are more informative for decision and policymaking, e.g., using absolute risks under alternative interventions.²⁰

What Do We Do in Practice?

The above is an attempt to sketch a unifying framework for assessing the fit of some meta-analysis models to data, and is motivated by the observations of Poole et al.⁴ We (Orestis and Tom) would be rather content were this sketch to hold to scrutiny, and to be fleshed out. Yet when it comes to practical meta-analysis, we do not envision routinely checking the goodness of various link functions or of alternative assumptions for random effects. After all, these explorations at best show which model appears to be supported by the data and do not constitute a procedure to recover the “correct” generative model that nature operates by. Such a task is never achievable with data alone.

In practice, we (Orestis and Tom) default to random effects models, because we find the equal effect assumption implausible. We choose effect measures heuristically, based on interpretability (for example, we have never analyzed clinical data using an entropy measure), statistical properties, and whether a parsimonious, mean-only, model fits data as well as a small set of alternative models (e.g., a model that includes the control rate as an explanatory variable^{16–18}). Furthermore, because most meta-analyses in health include only a small or modest number of studies, statistical tests between models would probably be underpowered, and noninformative. In our meta-analytic applications, we often use log odds ratios, but have used risk differences as well. We are not tempted to discover hidden laws of nature in the meta-epidemiological data that motivated the paper of Poole et al.⁴ Each meta-analysis is to us a distinct problem.

Some Further Thoughts on the Interpretation of “Meta-epidemiological” Observations

At best—at the very, very best—the simplistic meta-analysis models we use all the time are crude and hopefully serviceable descriptions of the observed studies. We find it strange to posit that studies included in a meta-analysis arise through an underlying data-generative process. (And how strange would a generative process be, were it to satisfy meta-analytic questions that typically arise in the minds of the meta-analysts after the studies have been completed!) No super-population of studies exists, from which we sample studies for our meta-analysis. And if global generative processes are unlikely to exist, it may be that the laws of nature that would govern them are not there to uncover. We have engaged in meta-epidemiological research ourselves, but came to appreciate that one should interpret findings rather narrowly, and refrain from bold statements and from extrapolating to each and every meta-analysis in hand. This exposition should not satisfy you, but it captures our thinking on the matter.

A Call for More Research on Describing and Exploring Heterogeneity in a Meta-analysis

We close with a call for more research on describing, measuring, and visualizing heterogeneity in hopes that researchers (and funders) will oblige. Poole et al. are right that methods that work across all effect measures are needed, but so much more should be done even conditional on an effect measure. Concepts related to “heterogeneity” (as used in

evidence synthesis in health) are those of “inequality” (as used in economics) and “diversity” (as used in ecology), and perhaps some cross-pollination would pay off. Finally, visualizations are useful complements to statistics,²³ and several nice examples exist.^{24,25}

POSTSCRIPT

So, does nature like multiplicative or additive effects? You should probably ask her every time you interact with her.

Acknowledgments

Supported by the Intramural Research Program of the National Cancer Institute.

The opinions expressed here are those of the authors and do not necessarily represent the official views of the US National Cancer Institute or the US Department of Health & Human Services.

Biographies

ORESTIS A. PANAGIOTOU is a researcher in the Biostatistics Branch of the Division of Cancer Epidemiology & Genetics (www.dceg.cancer.gov) at the US National Cancer Institute. He focuses on multidisciplinary applications of evidence synthesis in clinical, genetic, and molecular epidemiology.

THOMAS A. TRIKALINOS is Associate Professor of Health Services, Policy & Practice at Brown University, where he also directs the Center for Evidence-based Medicine (www.cebm.brown.edu). His work is on evidence synthesis (learning across multiple data sources) and evidence contextualization (using information to make optimal decisions).

REFERENCES

- Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med*. 2000; 19:1707–1728. [PubMed: 10861773]
- Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med*. 2002; 21:1575–1600. [PubMed: 12111921]
- Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol*. 2001; 54:1046–1055. [PubMed: 11576817]
- Poole C, Shrier I, VanderWeele T. Is the risk difference really a more heterogeneous measure? *Epidemiology*. 2015; 26:714–718. [PubMed: 26196684]
- Hardin, JW.; Hilbe, JM. *Generalized Linear Models and Extensions*. 3rd ed.. College Station, TX: Stata Press; 2012.
- Pregibon D. Goodness of link tests for generalized linear models. *J R Stat Soc Ser C (Appl Stat)*. 1980; 29:15–14.
- McCullagh, P.; Nelder, JA. *Monographs on Statistics and Applied Probability*. 2nd ed.. London; New York, NY: Chapman and Hall; 1989. *Generalized Linear Models*.
- Kovalchik SA, Varadhan R, Fetterman B, Poitras NE, Wacholder S, Katki HA. A general binomial regression model to estimate standardized risk differences from binary response data. *Stat Med*. 2013; 32:808–821. [PubMed: 22865328]
- Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol*. 1986; 123:174–184. [PubMed: 3509965]

10. Breslow NE. Generalized linear models: checking assumptions and strengthening conclusions. *Stat Appl.* 1996; 8:23–41.
11. Pan Z, Lin DY. Goodness-of-fit methods for generalized linear mixed models. *Biometrics.* 2005; 61:1000–1009. [PubMed: 16401273]
12. Raftery AE. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika.* 1996; 83:251–266.
13. Li, J. Choosing the proper link function for binary data. The University of Texas at Austin; 2014.
14. Ntzoufras I, Dellaportas P, Forster JJ. Bayesian variable and link determination for generalised linear models. *J Stat Plan Inference.* 2003; 111:165–180.
15. Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. *Stat Sinica.* 1996; 6:733–760.
16. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. *BMJ.* 1996; 313:735–738. [PubMed: 8819447]
17. Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat Med.* 1998; 17:1923–1942. [PubMed: 9777687]
18. McIntosh MW. The population risk as an explanatory variable in research synthesis of clinical trials. *Stat Med.* 1996; 15:1713–1728. [PubMed: 8870154]
19. Walter SD. Choice of effect measure for epidemiological data. *J Clin Epidemiol.* 2000; 53:931–939. [PubMed: 11004419]
20. Guyatt, G.; Rennie, D.; Meade, M.; Cook, D. American Medical Association. JAMA evidence. 3rd ed.. New York, NY: McGraw-Hill Education Medical; 2015. Users' Guides to the Medical Literature. A Manual for Evidence-Based Clinical Practice.
21. Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev.* 1987; 9:1–30. [PubMed: 3678409]
22. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. *Am J Epidemiol.* 1994; 140:290–296. [PubMed: 8030632]
23. Anzures-Cabrera J, Higgins JP. Graphical displays for meta-analysis: an overview with suggestions for practice. *Res Synth Methods.* 2010; 1:66–80. [PubMed: 26056093]
24. Jimenez FJ, Guallar E, Martin-Moreno JM. A graphical display useful for meta-analysis. *Eur J Public Health.* 1997; 7:101–105.
25. Olkin I, Dahabreh IJ, Trikalinos TA. GOSH – a graphical display of study heterogeneity. *Res Synth Methods.* 2012; 3:214–223. [PubMed: 26062164]