Published in final edited form as:

Genet Epidemiol. 2017 December ; 41(8): 876–886. doi:10.1002/gepi.22085.

On meta- and mega-analyses for gene–environment interactions

Jing Huang¹, Yulun Liu¹, Steve Vitale¹, Trevor M. Penning², Alexander S. Whitehead², Ian A. Blair², Anil Vachani³, Margie L. Clapper⁴, Joshua E. Muscat⁵, Philip Lazarus⁶, Paul Scheet⁷, Jason H. Moore¹, and Yong Chen¹

¹Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

²Department of Systems Pharmacology and Translational Therapeutics, Center of Excellence in Environmental Toxicology, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

³Pulmonary, Allergy, and Critical Care Division, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

⁴Cancer Prevention and Control Program, Fox Chase Cancer Center, Philadelphia, Pennsylvania, United States of America

⁵Department of Public Health Sciences, Penn State College of Medicine, Hershey, Pennsylvania, United States of America

⁶Department of Pharmaceutical Sciences, Washington State University, Spokane, Washington, United States of America

⁷Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America

Abstract

Gene-by-environment ($G \times E$) interactions are important in explaining the missing heritability and understanding the causation of complex diseases, but a single, moderately sized study often has limited statistical power to detect such interactions. With the increasing need for integrating data and reporting results from multiple collaborative studies or sites, debate over choice between mega- versus meta-analysis continues. In principle, data from different sites can be integrated at the individual level into a “mega” data set, which can be fit by a joint “mega-analysis.”

Alternatively, analyses can be done at each site, and results across sites can be combined through a

Correspondence: Jing Huang, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, PA 19104, United States of America. jing14@mail.med.upenn.edu; Paul Scheet, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, PA 19104, United States of America. pscheet@alum.wustl.edu; Jason H. Moore and Yong Chen, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, PA 19104, United States of America. jhmoore@mail.med.upenn.edu; ychen123@mail.med.upenn.edu.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ORCID

Jing Huang <http://orcid.org/0000-0002-6133-6988>

“meta-analysis” procedure without integrating individual level data across sites. Although mega-analysis has been advocated in several recent initiatives, meta-analysis has the advantages of simplicity and feasibility, and has recently led to several important findings in identifying main genetic effects. In this paper, we conducted empirical and simulation studies, using data from a $G \times E$ study of lung cancer, to compare the mega- and meta-analyses in four commonly used $G \times E$ analyses under the scenario that the number of studies is small and sample sizes of individual studies are relatively large. We compared the two data integration approaches in the context of fixed effect models and random effects models separately. Our investigations provide valuable insights in understanding the differences between mega- and meta-analyses in practice of combining small number of studies in identifying $G \times E$ interactions.

Keywords

fixed effect model; gene–environment interaction; mega-analysis; meta-analysis; random-effects model

1 INTRODUCTION

Gene-by-Environment ($G \times E$) interactions play a vital role in explaining the missing heritability and understanding the causation of complex diseases (Brockton, Little, Sharp, & Cotton, 2000; Doll & Peto, 1981; Khoury, Adams, & Flanders, 1988; Khoury, Beaty, & Hwang, 1995; Le Marchand & Wilkens, 2008; Manolio et al., 2009; Thomas, 2010). It typically requires larger sample size to identify such interaction effects than those needed to identify genetic main effects of comparable magnitude (Smith & Day, 1984; Thomas, 2010). In order to improve the statistical power of detecting $G \times E$ interactions, it is crucial to combine data and findings from multiple studies. To this end, large research consortia are increasingly available, including CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology; Psaty et al. 2009), ENGAGE (European Network of Genomic and Genetic Epidemiology; Codd et al. 2013), PGC (Psychiatric Genomics Consortium), ICBP (International Consortium of Blood Pressure; Ehret et al. 2011), and CArE (Candidate Gene Association Resource; Musunuru et al. 2010).

Despite the increasing availability of data from multiple studies, there is still a lack of guidelines on whether mega- or meta-analysis should be used to combine data from different studies, and debate on the choice between them continues (Steinberg et al., 1997; Stewart & Tierney, 2002; Sung et al., 2014). In principle, it may seem ideal to pool individual level data from multiple studies into a “mega” data set and conduct a single analysis. However, in the context of identifying $G \times E$ interactions, it is inconvenient, time-consuming, and sometimes logistically challenging to pool genetic data from different studies, due to diverse terms of informed consents, redundant and possibly nonuniform quality control protocols between research institutions, and privacy preservation issues. Alternatively, meta-analysis is much more practical. Analyses are conducted within each site, and results across sites (effect sizes and standard errors) are combined without sharing individual level data. In fact, meta-analysis has led to important findings in identifying genes with main genetic effects that

explain heritability in heights, BMI, and their variability in recent research (Allen et al., 2010; Speliotes et al., 2010; Yang et al., 2012).

Motivated by the ongoing debates between the use of mega- or meta-analysis, investigators have recently conducted both theoretical and empirical investigations on the relative performances comparing the two analyses (Lin & Zeng, 2010a, 2010b; Sung et al., 2014; Zeng & Lin, 2015). The comparisons between the two approaches have been conducted under the context of fixed-effect models and random-effects models separately. Specifically, fixed-effect mega-analysis assumes a common effect size of $G \times E$ interaction across all studies and uses the pooled individual level data to estimate this common effect. Fixed effect meta-analysis makes the same common effect assumption and combines the effect size estimates from different studies using an inverse-variance weighting procedure. An empirical comparison between fixed-effect mega- and meta-analyses of $G \times E$ interaction has been conducted by Sung et al. (2014), where they used data from four studies of cardiovascular disease to compare the effect sizes, standard errors, and P -values from both analyses. They concluded that both analyses yield comparable results for studying $G \times E$ interaction in fixed-effect models. Their conclusions were consistent with the theoretical findings by Lin and Zeng (2010b), where they showed that the estimator of common effect from fixed-effect meta-analysis follows the same limiting distribution as the estimator from fixed-effect mega-analysis. In other words, there is no efficiency loss (no power loss) to use fixed-effect meta-analysis, compared to fixed-effect mega-analysis, when the sample sizes of individual studies are sufficiently large.

In addition to fixed-effect models, Zeng and Lin (2015) also conducted theoretical investigations to compare random-effects mega- and meta-analyses. Specifically, random-effects mega-analysis allows each site to have a unique site-specific effect and uses random-effects models to account for the heterogeneity. Random effects meta-analysis makes the same assumption of site-specific effects and combines the distinct estimates of effect sizes from different sites using a normal distribution. In some applications, the assumption of site-specific effects can be more appropriate than the assumption of a common effect across studies due to the heterogeneity of population across sites, differences in study design, and differential exposures from unmeasured covariates or even confounders. Zeng and Lin (2015) showed that, when both the number of studies and sample sizes of individual studies are large, the random-effects meta-analysis estimator is at least as efficient as the random-effects mega-analysis estimator. In another scenario where the number of study is small and the sample size of each study is large, the two estimators follow two distinct mixtures of normal distributions, and the theoretical comparison between them is challenging because the mixing probabilities of the two mixture distributions are model dependent.

Motivated from a recent study of identifying $G \times E$ interactions for lung cancer, where the sequenced single-nucleotide polymorphisms (SNPs) are selected based on a specific SNP CHIP design such that the genetics data of patients from only four sites are available, we are interested in comparing the mega- and meta-analyses in studies of $G \times E$ interactions under the scenario that the number of study is small and the sample sizes of individual studies can be relatively large. Along the line of the investigations by Lin and Zeng (2010a, 2010b), Sung et al. (2014), Zeng and Lin (2015), our unique contributions are as follows.

1. By comparing results from fixed-effect mega- and meta- analyses, we investigate the sample size needed for fixed-effect meta-analysis to yield comparable results with mega-analysis. Such investigation can provide practical guidance for choosing between mega- and meta-analyses in studying $G \times E$ interactions when fixed-effect models are used.
2. In the theoretical exploration by Zeng and Lin (2015), they concluded that the comparison of random-effects mega- and meta-analyses is model dependent. In this paper, we consider commonly used models for studying $G \times E$ interactions, and empirically compare the results from random-effects mega- and meta-analyses when these models are used.
3. Existing literature has shown that fixed-effect mega- and meta-analyses have similar results under the assumption of no between-study heterogeneity (Lin & Zeng, 2010b). In this paper, we empirically showed that the results from two analyses are also similar as long as the heterogeneity is not substantial.
4. From a practical perspective, we compare the sets of top SNPs identified by mega- and meta-analyses in detection of novel $G \times E$ interactions for both fixed-effect and random-effects models. We find the proportion of the overlap of sets of detected SNPs increases as sample size of each study increases for the fixed-effect model, but such a trend is not observed for the random-effects model.

In the rest of the paper, we first briefly review the fixed-effect and random-effects model in a generic generalized linear model setting and review the results of Lin and Zeng (2010b) and Zeng and Lin (2015), then we discuss four commonly used analyses in identifying $G \times E$ interactions. After that we show the results of our investigations and provide a brief discussion.

2 METHOD

2.1 Fixed-effect model

Consider integrating data from K studies with n_k participants in the k th study. The goal is to quantify the effects of predictors ($\mathbf{X}$) on an response (Y) adjusting for covariates $\mathbf{Z}$. Let subscript ki denote the i th subject in the k th study, $i = 1, 2, \dots, n_k$ and $k = 1, 2, \dots, K$. Under the assumption that the effects of $\mathbf{X}$ on Y are common across K study populations, a fixed-effect model in the generalized linear model setting can be formulated as

$$g(\mu_{ki}) = \alpha_k + \beta \mathbf{X}_{ki} + \gamma_k \mathbf{Z}_{ki}, \text{ with}$$

$$\text{var}(Y_{ki}) = \phi_k V(\mu_{ki}) \text{ and} \quad (1)$$

$$\boldsymbol{\eta}_k \equiv (\alpha_k, \boldsymbol{\gamma}_k, \phi_k),$$

where $\mu_{ki} = E(Y_{ki})$, $g(\cdot)$ is a link function, for example, logit function in logistic regression, $V(\mu_{ki})$ is a function of μ_{ki} , and ϕ_k is the dispersion parameter.

Notably, the effects of interest, β , which are the effects of X on Y , adjusting for covariates Z , is assumed to be common across studies, while the other parameters, the nuisance parameters η_k , are allowed to be study-specific. Under model (1), testing for an effect of X on Y is equivalent to test

$$H_0: \beta = 0.$$

When data of individual participants are available from all the studies, the fixed-effect meta-analysis estimator of β can be obtained by maximizing the joint likelihood function constructed based on Equation (1) using the individual level data. When only the effect size and variance estimates of β are available from each study, the fixed-effect meta-analysis estimator of β can be obtained by combining the estimates using an inverse-variance weighting procedure (Hedges & Olkin, 1985). When the sample sizes of individual studies are large, both fixed-effect meta- and meta-analysis estimators converge to the same limiting distribution (Lin & Zeng, 2010b). In other words, inferences based on the fixed-effect meta- and meta-analyses are comparable when the sample sizes of individual studies are large.

2.2 Random-effects model

The random-effects models assume the effects of interest, that is, the effects of X on Y adjusting for covariates Z , are study-specific and follow a normal distribution to allow for between-study heterogeneity, such that

$$g(\mu_{ki}) = \alpha_k + \beta_k X_{ki} + \gamma_k Z_{ki}, \text{ with}$$

$$\beta_k \sim N(\beta, \tau^2), \quad (2)$$

$$\text{var}(Y_{ki}) = \phi_k V(\mu_{ki}), \text{ and}$$

$$\eta_k \equiv (\alpha_k, \gamma_k, \phi_k, \tau^2),$$

for $i = 1, 2, \dots, n_k$ and $k = 1, 2, \dots, K$. Different from the fixed-effect model (Equation (1)), the regression coefficients β_k in Equation (2) is the study-specific effects of X on Y , after adjusting for the covariates Z . Under model (2), testing for the effects of X is equivalent to testing

$$H_0: \beta = 0.$$

Similarly, when data of individual participants are available, the random-effects mega-analysis estimator of the overall effect β can be obtained by maximizing the joint likelihood function constructed based on Equation (2) using the individual level data from all K studies. When only the effect size and variance estimates of β_k are available from each study, the random-effects meta-analysis assumes the estimates from each study, $\hat{\beta}_k$, follow a distribution with mean β and variance consisting of both within-study variation and between-study heterogeneity (DerSimonian & Laird, 1986). The estimation of β is obtained by first estimating the between-study variance of β_k (i.e. the variance of random effects) and then estimating the overall effect β and the standard error.

Zeng and Lin (2015) compared the random-effects mega- and meta-analysis estimators under two circumstances: (1) both sample sizes of individual studies and the number of studies are very large, and (2) the sample sizes of individual studies are large but the number of studies is small. They proved that, under the first circumstance, variance of the random-effects meta-analysis estimator is always less than or equal to that of the random-effects mega-analysis estimator, suggesting the random-effects meta-analysis is better than the random-effects mega-analysis in combining data from a large number of large studies. Under the second circumstance, they found the comparison between random-effects mega- and meta-analyses is difficult to be generalized, because the limiting distributions of both estimators are model dependent.

In this manuscript, we conduct empirical comparisons between mega- and meta-analyses in both fixed-effect and random-effects models. Specifically, we investigate the second scenario in Zeng and Lin (2015) that the number of studies is small and the sample sizes of individual studies can be large, and we conduct the comparison for four types of analyses that are commonly used in $G \times E$ studies. Our investigations are motivated from a recent genetic study of lung cancer, in which the SNPs sequenced are selected based on a specific SNP CHIP design (e.g., an SNP CHIP designed to target particular genetic locations that are potentially associated with lung cancer), such that the available number of studies containing the SNP data are limited. Our findings can provide practical guidance on the choice between mega- and meta-analyses in these situations.

3 ANALYSES FOR IDENTIFICATION OF $G \times E$ INTERACTIONS

In the analysis of $G \times E$ interactions, commonly there are three observed variables: the response trait (Y), environmental variable (E), and genetic variable (G). Here we discuss four types of analyses considered in Sung et al. (2014): the genetic marginal effect analysis, standard analysis of $G \times E$ interaction, genetic main effect analysis in the presence of $G \times E$ interaction, and joint analysis of both genetic main effect and $G \times E$ interaction effect.

In the following subsections, we show that the fixed-effect or random-effects models for each of these four analyses can be formulated using Equation (1) or (2), respectively. Hence, inferences based on the fixed-effect mega- and meta-analyses of $G \times E$ interactions in these analyses are asymptotically equivalent, but the comparability between the estimators of the random-effects mega- and meta-analyses needs to be investigated with respect to each analysis.

3.1 Analysis 1: Genetic marginal effect analysis

We first compare the mega- and meta-analyses using the genetic marginal effect analysis, which is most commonly used in GWAS. It tests for the overall effect of G on Y using the regression model

$$g\{E(Y)\} = \delta_0 + \delta_e E + \delta_g G, \quad (3)$$

where Y , E , and G are the response variable, environment variable, and genetic variable, respectively, and δ_0 , δ_e , and δ_g denote the intercept, environmental marginal effect, and genetic marginal effect, respectively. In this analysis, the parameter of interest is δ_g . A fixed-effect model for K studies can be formulated using Equation (1) to assume δ_g is a common effect for all the studies. A random-effects model can be formulated using Equation (2) to allow the study-specific genetic marginal effects, that is, the k th study has a unique δ_{gk} , $k = 1, \dots, K$, and they follow a normal distribution $\delta_{gk} \sim \mathcal{N}(\delta_g, \tau^2)$. The hypothesis is to test $H_0 : \delta_g = 0$.

Similarly, the models for the three additional analyses described next can also be formulated using Equation (1) or (2) when the fixed-effect or random-effects model is considered, respectively.

3.2 Analysis 2: Standard interaction effect analysis

A commonly used regression model to identify $G \times E$ interaction is

$$g\{E(Y)\} = \gamma_0 + \gamma_e E + \gamma_g G + \gamma_I G \times E, \quad (4)$$

where γ_0 , γ_e , γ_g , and γ_I are the intercept, environmental main effect, genetic main effects, and multiplicative $G \times E$ interaction effect, respectively. In the interaction effect analysis, the parameter of interest is γ_I and the null hypothesis is $H_0 : \gamma_I = 0$.

3.3 Analysis 3: Genetic main effect analysis in the presence of $G \times E$ interaction

The analysis of genetic main effect in the presence of $G \times E$ interaction uses the same model as shown in Equation (4), and tests for the main effect of G on Y adjusting for the multiplicative interaction effect of G and E on Y . The parameter of interest is γ_g and the null hypothesis of no genetic main effect after adjusting for the interaction effect is $H_0 : \gamma_g = 0$.

3.4 Analysis 4: Joint analysis of both genetic main effect and $G \times E$ interaction

The joint analysis of both genetic main effect and $G \times E$ interaction effect also uses the same model as shown in Equation (4), and tests for both genetic main effect and the interaction effect to be 0, that is, $H_0 : \gamma_g = \gamma_I = 0$. The test has 2 degrees of freedom.

In the next section, we first describe the motivating study of lung cancer, in which genetics data are available for participants from four hospitals in Pennsylvania. We show the comparisons of mega- and meta-analyses with increasing sample sizes of individual studies

(increasing from 10% to 100% of the data) to investigate the trend of the results. We present the comparison results for fixed-effect model and random-effects model separately.

4 EMPIRICAL STUDIES

Lung cancer is the leading cause of cancer death for men and women in the United States, and smoking has been understood as a major risk factor. As the majority (up to 90%) of lung cancers occur in smokers, but only a minority (~10%) of smokers get the disease (Cassidy, Duffy, Myles, Liloglou, & Field, 2007), it is likely that significant gene/phenotype/environmental interaction exist (Ihsan et al., 2011). In a study of $G \times E$ interactions in lung cancer conducted in the Pennsylvania area, a total of 1,918 subjects were recruited from four different sites: Fox Chase Cancer Center, Hershey Medical Center, Temple University Hospital, and the University of Pennsylvania Medical Center. Genotyping was performed for 1,610 of the 1,918 subjects using a targeted genotyping approach of candidate genes (Baldwin et al., 2013). Numbers of cases and control with genotyping data from each site are summarized in Table 1.

Here, we conduct both mega- and meta-analyses using the four common analyses aforementioned to assess the $G \times E$ interactions between smoking and genetic markers. We group the subjects by recruitment site to create four distinct study groups. The environmental variable is average quantity of cigarettes smoked per day. Subjects without smoking assessment were excluded. The analyses are conducted for 1,206 subjects and a total of 11,345 SNPs. In order to compare the approaches in different sample sizes and to observe the trend of the results, we take a sequence of subsamples with increasing sample size by randomly sampling 10%, 25%, 75%, and 100% of subjects within each of the four sites, and conduct the analyses for each sample. This strategy of subsampling allows us to gain insights into such a comparison at larger sample sizes than actually observed by observing the trend and stability in relative behaviors of these approaches. We conduct the comparisons in the context of fixed-effect and random-effects models separately.

4.1 Comparison of fixed-effect mega- versus meta-analysis

To compare the fixed-effect mega- and meta-analyses, we evaluate the agreement of the results obtained from the two approaches by calculating the intraclass correlation coefficient (ICC) between the two test statistics and between the effect size estimates. The ICC is a better measure of agreement than the Pearson correlation due to that it is invariant to linear transformations of the two variables being compared. If one of the variables is a linear transformation of the other variable, for example, $X = 2Y$, the Pearson correlation may still show a perfect correlation. This property is not desired if we want to evaluate the equivalence of the two approaches. On contrary, ICC will show perfect correlation, only when the two variables are the same, that is, $X = Y$.

According to the theoretical results of Lin and Zeng (2010b), the fixed-effect meta-analysis is equivalent to the fixed-effect mega-analysis for most regression models when the sample sizes of individual studies are large. The empirical results of our study confirm this conclusion in the four analyses of $G \times E$ interactions. The ICC between both test statistics and point estimates obtained through mega- and meta-analyses converges to 1 as sample

sizes of individual studies increase, which suggests that inferences based on the fixed-effect mega- and meta-analyses become equivalent when the sample sizes of individual studies become larger. We also find that the sample sizes of individual studies do not need to be very large for the two test statistics to have high agreement in the settings we explored. In our analyses, the ICC values of the two test statistics are greater than 0.8 when only 25% of the data are used (46, 106, 30, 222 patients in four sites, and 404 in total) for all four analyses. We also note that the fixed-effect mega- and meta-analyses have higher concordance for SNPs with large test statistics, which suggests that the two methods are consistent in identifying markers with $G \times E$ interactions. Similar results are found for all the four analyses of $G \times E$ interaction. Details are shown in supplementary information Appendix 1.

We also construct Venn diagrams to compare the two sets of top markers identified by fixed-effect mega- and meta-analyses. We used P -value = 0.01 as a cutoff for the significant level. We do not use the Bonferroni correction or false discovery rate to adjust for the multiple comparison issue here, first because either method will yield only few (< 10) markers to pass the threshold which made the comparison less intuitive, and second because the purpose of this real data application is to compare two statistical methods, rather than to find new markers with significant interactions. Figure 1 shows the Venn diagrams plotted using the number of markers identified by fixed-effect meta- and mega-analyses using 25% data (left panel) and 100% data (right panel). Each row represents one of the four analyses of $G \times E$ interaction (from top to bottom: the genetic marginal effect analysis, standard interaction effect analysis, genetic main effect analysis in the presence of $G \times E$ interaction, and joint analysis of both genetic main effect and $G \times E$ interaction). We observe that, as sample sizes of individual studies increase, the total number of markers identified by either fixed-effect mega- or meta-analysis increases because increased sample size results in better sensitivity for small interaction effects. More importantly, the Jaccard Index (number of overlapped markers divided by totally number of marker identified) indicates the similarity between the two sets of markers identified by fixed-effect meta- and mega-analysis increases as sample size of each study increases. Such a finding confirms the agreement between the two methods when sample sizes of individual studies are large.

We also check the between-study heterogeneity of the effects of interest for every SNPs, using I^2 in each analysis. We note that more than 65% of SNPs have $I^2 < 25\%$, suggesting that the fixed-effect model is more appropriate for most of the SNPs. It may be due to the fact that all four sites are in Pennsylvania thus the populations are relatively homogenous. We also conduct the comparison of fixed-effect mega- and meta-analyses in two different sets of SNPs, one with $I^2 < 25\%$ and the other with $I^2 > 25\%$. The conclusions are similar. The agreement of results between fixed-effect mega- and meta-analyses increases as sample sizes of individual studies increase, suggesting that even when the common effect assumption may not hold, the fixed-effect mega- and meta-analyses still yield comparable results when number of studies is small and sample sizes of individual studies are relatively large.

4.2 Comparison of random-effects mega- versus meta-analysis

We also compare the estimates of $G \times E$ interactions from random-effects mega- and meta-analyses. According to Zeng and Lin (2015), the random-effects meta-analysis is at least as efficient as the mega-analysis when both sample sizes of individual studies and number of studies are large, in that the variance of mega-analysis estimator is always greater than or equal to the variance of meta-analysis estimator. Their work also suggested that the comparison between the random-effects mega- and meta-analyses is difficult to be generalized when the number of studies is small and sample sizes of individual studies are large, because the limiting distribution of the two estimators are model depended in this scenario.

In our investigation, we compare the two approaches in the second scenario in the four commonly used models of $G \times E$ interactions. We compare the relative efficiency of random-effects mega- and meta-analysis estimators using the ratio of empirical variances of the estimated regression coefficients (for the 11,345 SNPs) obtained from both approaches. This quantity is slightly different with the conventional relative efficiency for two estimators, because the empirical variance of the estimated regression coefficients consists of two variance components: the variance of the true effect sizes across the SNPs (the true effect sizes of the SNPs are different because it is a real rather than simulated data set), and the sum of the empirical variance of the estimates for all SNPs. The first component is a fixed value that is intrinsic and decided by the true associations between the SNPs and the disease condition, and the second component quantifies the variance of the estimator. Thus, the comparison of such an empirical variance still provides insights on the actual relative efficiency of the random-effects meta- and mega-analysis estimators of $G \times E$ interactions. We do not compare the test statistics from random-effects mega- and meta-analyses because the two test statistics converge to different distributions.

In Figure 2, we show the estimated effect sizes for the genetic marginal effect analysis from the random-effects mega-analysis(y-axis) versus the random-effects meta-analysis(x-axis), as sample sizes of each studies increase from 10% to 100% data. We observe the ICC does not converge to 1 in this analysis. The same phenomenon is also found in the other three analyses (shown in supplementary information Appendix 2). It suggests that the two approaches are not equivalent in the scenario we considered, even when sample sizes of individual studies are relatively large. Interestingly, we observe the “relative efficiency” between the estimators of random-effects meta- and mega-analyses is >1 when 100% data are used, and this phenomenon is consistently observed in all the four analyses of $G \times E$ interaction (shown in supplementary information Appendix 2). It suggests that the random-effects meta-analysis is more efficient (i.e., the variance of the estimator is smaller) than the random-effects mega-analysis when number of studies is small and sample sizes of individual studies are relatively large.

To validate our conclusion, we conduct simulation studies guided by the lung cancer data to further compare the two approaches. Using the genetic marginal effect model (Equation (3)), we simulate a data set with four studies, each with 1,000 subjects under the null hypothesis. The environmental variable (number of cigarettes per day) is generated from four truncated normal distributions for subjects in different studies, with means and variances calculated

from the lung cancer data. Genetic variables are simulated with minor allele frequencies randomly selected from uniform distribution $U(0.2, 0.4)$. The intercept and environmental marginal effect for each study are simulated from two normal distributions, with means and variances calculated using the lung cancer data, to allow between-study heterogeneity. The genetic marginal effects for each study are simulated from a normal distribution, with mean equals 0 and variance calculated from the lung cancer data. We simulate the binary phenotype variable based on Equation (3). The simulation is conducted for 10,000 replicates. We use qqplot to compare the test statistics obtained from the random-effects mega- and meta-analyses with χ_1^2 distribution, and find that both test statistics deviate from χ_1^2 and can result in hypothesis tests with inflated Type I errors, shown in Figure 3(a) and (b). Such inflation is due to the underestimation of the variance of random effects when the number of studies is small (Kosmidis, Guolo, & Varin, 2017). In our simulation, we also compare the test statistics from the two approaches, shown in Figure 3(c). We observe that the test statistic from random-effects mega-analysis is more dispersed than that from the random-effect meta-analysis, suggesting random-effects meta-analysis has smaller inflation than random-effects mega-analysis in identification of $G \times E$ interactions when number of studies is small and sample sizes of individual studies are relatively large.

We also construct the Venn diagrams to compare the sets of top markers identified by both random-effects mega- and meta-analyses, shown in Figure 4. We observe that for both approaches, the total number of identified markers increases as sample sizes of individual studies get larger, which is consistent with our finding in the fixed-effect model, and suggests that larger sample sizes of individual studies also provide more power to detect smaller $G \times E$ interactions in random-effects model. We do not observe an obvious increasing trend in the total proportion of overlapped markers identified by the two approaches as sample sizes of individual studies increase. It is due to the fact that the two estimators converge to different distributions.

5 DISCUSSION

In order to increase the effective sample size and power of detecting genetic effects, it has been increasingly popular to report results using combined data/statistics from multiple studies and even among large consortia. This holds for genetic interactions, as well, including $G \times E$ interactions, although these have been less studied. Mega-analysis that analyzes pooled individual level data faces many challenges in practice, such as the storage and management between institutes, the difficulties of integrating very large genotype and phenotypic data sets from different studies, and the issue of data sharing. Instead, meta-analysis that combines the summary statistics from multiple studies has several advantages: (1) data can be analyzed within each institute, (2) each study can adjust for study-specific covariates and local population substructure, and (3) inferences on $G \times E$ interactions are made without sharing individual level data.

In this paper, we conducted empirical investigations to compare mega- and meta-analyses in analyses of $G \times E$ interaction under the context of fixed-effect model and random-effects model, respectively. Motivated from a recent lung cancer study, which uses a specific SNP CHIP design such that the number of studies is limited, our investigation focused on the

scenario when number of studies is small and the sample sizes of individual studies are relatively large. Using a sequence of increasing percentage of subsample from the lung cancer study, we attempted to gain insights into such a comparison at larger sample sizes than actually observed by getting a handle on the trend and stability in relative behaviors of these approaches. We used the ICC to measure the equivalence between the mega- and meta-analyses. We found that the fixed-effect mega- and meta-analyses yield similar effect size estimates and test statistics as the sample sizes of individual studies increase. This conclusion is concordant with the main findings in Lin and Zeng (2010b) and Sung et al. (2014). We also observed that fixed-effect meta-analysis can yield comparable results to fixed-effect mega-analysis even when the sample sizes of individual studies are moderate. This is important because it justifies the use of meta-analysis as a surrogate for mega-analysis when a fixed-effect model is assumed, where the mega-analysis requires pooling individual level data together, facing the practical challenges in data sharing and others.

Of note, when a random-effects model was adapted, we found that the estimator of random-effects meta-analysis is more efficient than that of the random-effects mega-analysis in analyzing $G \times E$ interactions in our scenario. We observed that both approaches yield inflated Type I errors when the number of studies is small, but the inflation of random-effects meta-analysis is smaller than that of the random-effects mega-analysis in our analyses of $G \times E$ interactions, suggesting the random-effects meta-analysis is better than random-effects mega-analysis when number of studies is small.

From a practical perspective, we also compared the overlapped top markers identified by mega- and meta-analysis. When a fixed-effect model was used, the proportion of overlapped top markers increased as sample sizes of individual studies increased, but the trend was not observed for the random-effects models. This empirical finding highlights the research gap that more considerations on choice between mega- and meta-analysis are needed when a random-effects model is assumed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research reported in this publication was supported in part by the National Institutes of Health under award numbers R01 AI130460 (for JH and YC), R01 AI116794 (for JHM and YC), R01 LM009012 (for YC), R01 HG005859 (for PS and YL), P42 ES023720 (for IAB), and P30 ES013508 (for ASW, AV, IAB, JHM and TMP). This research was also supported by the Pennsylvania Department of Health under award number PA 4100038714 (for ASW, AV, IAB, JHM, MLC, PL and TMP).

References

- Allen HL, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Hirschhorn JN. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*. 2010; 467(7317):832–838. [PubMed: 20881960]
- Baldwin DA, Sarnowski CP, Reddy SA, Blair IA, Clapper M, Lazarus P, Whitehead AS. Development of a genotyping microarray for studying the role of gene–environment interactions in risk for lung cancer. *Journal of Biomolecular Techniques*. 2013; 24(4):198–217. [PubMed: 24294113]

- Brockton N, Little J, Sharp L, Cotton S. N-acetyltransferase polymorphisms and colorectal cancer: A huge review. *American Journal of Epidemiology*. 2000; 151(9):846–861. [PubMed: 10791558]
- Cassidy A, Duffy SW, Myles JP, Liloglou T, Field JK. Lung cancer risk prediction: A tool for early detection. *International Journal of Cancer*. 2007; 120(1):1–6.
- Codd V, Nelson CP, Albrecht E, Mangino M, Deelen J, Buxton JL, Samani NJ. Identification of seven loci affecting mean telomere length and their association with disease. *Nature Genetics*. 2013; 45(4):422–427. [PubMed: 23535734]
- DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials*. 1986; 7(3):177–188. [PubMed: 3802833]
- Doll R, Peto R. The causes of cancer: Quantitative estimates of avoidable risks of cancer in the United States today. *Journal of the National Cancer Institute*. 1981; 66(6):1192–1308.
- Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, Johnson T. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*. 2011; 478(7367):103–109. [PubMed: 21909115]
- Hedges, L., Olkin, I. *Statistical methods for meta-analysis*. Orlando, FL: Academic; 1985.
- Ihsan R, Chauhan PS, Mishra AK, Yadav DS, Kaushal M, Sharma JD, Saxena S. Multiple analytical approaches reveal distinct gene–environment interactions in smokers and non smokers in lung cancer. *PloS One*. 2011; 6(12):e29431. [PubMed: 22206016]
- Khoury MJ, Adams M Jr, Flanders WD. An epidemiologic approach to ecogenetics. *American Journal of Human Genetics*. 1988; 42(1):89–95. [PubMed: 3337114]
- Khoury MJ, Beaty TH, Hwang SJ. Detection of genotype-environment interaction in case-control studies of birth defects: How big a sample size? *Teratology*. 1995; 51(5):336–343. [PubMed: 7482355]
- Kosmidis I, Guolo A, Varin C. Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression. *Biometrika*. 2017; 104(2):489–496.
- Le Marchand L, Wilkens LR. Design considerations for genomic association studies: Importance of gene–environment interactions. *Cancer Epidemiology Biomarkers & Prevention*. 2008; 17(2):263–267.
- Lin D, Zeng D. Meta-analysis of genome-wide association studies: No efficiency gain in using individual participant data. *Genetic Epidemiology*. 2010a; 34(1):60–66. [PubMed: 19847795]
- Lin D, Zeng D. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*. 2010b; 97(2):321–332. [PubMed: 23049122]
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, Visscher PM. Finding the missing heritability of complex diseases. *Nature*. 2009; 461(7265):747–753. [PubMed: 19812666]
- Musunuru K, Lettre G, Young T, Farlow DN, Pirruccello JP, Ejebe KG, for the NHLBI Candidate Gene Association Resource. Candidate gene association resource (care). *Circulation: Cardiovascular Genetics*. 2010; 3(3):267–275. [PubMed: 20400780]
- Psaty BM, O'donnell CJ, Gudnason V, Lunetta KL, Folsom AR, Rotter JJ, CHARGE Consortium. Cohorts for heart and aging research in genomic epidemiology (charge) consortium. *Circulation: Cardiovascular Genetics*. 2009; 2(1):73–80. [PubMed: 20031568]
- Smith P, Day N. The design of case-control studies: The influence of confounding and interaction effects. *International Journal of Epidemiology*. 1984; 13(3):356–365. [PubMed: 6386716]
- Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Loos RJ. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*. 2010; 42(11):937–948. [PubMed: 20935630]
- Steinberg K, Smith S, Stroup D, Olkin I, Lee N, Williamson G, Thacker S. Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. *American Journal of Epidemiology*. 1997; 145(10):917–925. [PubMed: 9149663]
- Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & The Health Professions*. 2002; 25(1):76–97. [PubMed: 11868447]

- Sung YJ, Schwander K, Arnett DK, Kardia SL, Rankinen T, Bouchard C, Rao DC. An empirical comparison of meta-analysis and mega-analysis of individual participant data for identifying gene–environment interactions. *Genetic Epidemiology*. 2014; 38(4):369–378. [PubMed: 24719363]
- Thomas D. Gene–environment-wide association studies: Emerging approaches. *Nature Reviews Genetics*. 2010; 11(4):259–272.
- Yang J, Loos RJ, Powell JE, Medland SE, Speliotes EK, Chasman DI, Visscher PM. Fto genotype is associated with phenotypic variability of body mass index. *Nature*. 2012; 490(7419):267. [PubMed: 22982992]
- Zeng D, Lin D. On random-effects meta-analysis. *Biometrika*. 2015; 102(2):281–294. [PubMed: 26688589]

FIGURE 1.

Venn diagram showing proportion of overlapped top markers identified with P -value ≤ 0.01 by fixed-effect mega- and meta-analyses, measured by Jaccard Index. With sample sizes of individual studies increase, the Jaccard Index increases. Left panels are results using 25% data, right panels are results using 100% data. Each row represents different analyses of $G \times E$ interactions (from top to bottom: the genetic marginal effect analysis, the standard interaction effect analysis, the genetic main effect analysis in the presence of $G \times E$ interaction, and the joint analysis of both genetic main effect and $G \times E$ interaction)

FIGURE 2.

Plot of estimated effect size for the genetic marginal effect analysis from random-effects mega-analysis (y-axis) versus meta-analysis (x-axis)

FIGURE 3.

(a) qqplot of χ^2_1 distribution versus test statistic from random-effects meta-analysis of genetic marginal effect analysis using simulated data set with four studies, each with 1,000 subjects, over 10,000 replicates; (b) qqplot of χ^2_1 distribution versus test statistic from random-effects mega-analysis of genetic marginal effect analysis using simulated data set with four studies, each with 1,000 subjects, over 10,000 replicates; (c) qqplot of test statistic from random-effects mega-analysis (y -axis) versus test statistic from random-effects meta-analysis (x -axis) of the genetic marginal effect analysis using simulated data set with four studies, each with 1,000 subjects, over 10,000 replicates

FIGURE 4.

Venn diagram showing proportion of overlapped top markers identified with P -value ≤ 0.01 by random-effects mega- and meta-analyses, measured by Jaccard Index. With sample sizes of individual studies increase, the Jaccard Index does not increase. Left panels are results using 25% data, right panels are results using 100% data. Each row represents different analyses of $G \times E$ interaction (from top to bottom: the genetic marginal effect analysis, the

standard interaction effect analysis, the genetic main effect analysis in the presence of $G \times E$ interaction, and the joint analysis of both genetic main effect and $G \times E$ interaction)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE 1

Number of cases and controls with genotyping data from four sites of the $G \times E$ interaction study of lung cancer

Site	Cases	Controls	Total
Fox Chase Cancer Center	113	71	184
Hershey Medical Center	174	247	421
Temple University Hospital	61	57	118
University of Pennsylvania Medical Center	519	368	887
Total	867	743	1,610