Previous' lecture

- Single variant association
- Use genome-wide SNPs to account for confounding (population substructure)
- Estimation of effect size and winner's curse

Meta-Analysis

Today's outline

- *P*-value based methods.
- Fixed effect model.
- Meta vs. pooled analysis.
- Random effects model.
- Meta vs. pooled analysis
- New random effects analysis

Meta-Analysis

- Single study is under-powered because the effect of common variant is very modest (OR ≤ 1.4)
- Meta-analysis is an effective way to combine data from multiple independent studies

Meta-Analysis Methods

- *P*-value based
- Regression coefficient based
  - Fixed effects model
  - Random effects model

P-value Based

- Conduct meta-analysis using *p*-values
- Simple and widely used
- Fisher and *Z*-score based methods

P-value Based

- K studies

$$T_{\text{Fisher}} = \sum_{k=1}^{K} -2\log(p_k) \sim \chi^2_{2K}$$

- Simple and works well
- Direction of effect is not considered

P-value Based

- Stouffer's Z-score (one-sided right-tailed)

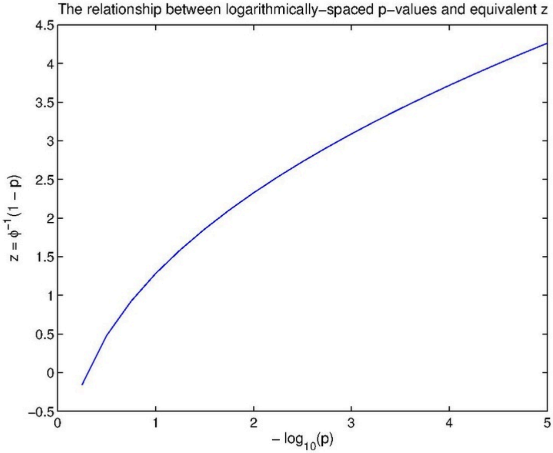
$$Z_k = \Phi(1 - p_k)$$

- Φ is the standard cumulative normal distribution function

$$Z = \frac{\sum_{k=1}^{K} Z_k}{\sqrt{K}}$$

- Z follows the standard normal distribution.

Fisher versus Z-score



- These two are not perfectly linear, but they follow a highly linear relationship over the range of Z values most observed ($Z \geq 1$).

P-value Based

- Z-score can incorporate direction of effects and weighting scheme
  - β_K : effect for study k
  - $w_k = \sqrt{n_k}$

$$Z_k = \Phi(1 - p_k/2) \cdot \operatorname{sign}(\beta_k)$$

$$Z = \frac{\sum_{k=1}^{K} w_k Z_k}{\sqrt{\sum_{k=1}^{K} w_k^2}} \sim N(0, 1)$$

P-value Based

- Easy to use
- *Z*-score is perhaps more popular as it works well to find a consistent signal from studies
- Cannot estimate the effect size

Regression Coefficient Based

- For each study $k = 1, \ldots, K$

$$g\{E(Y_{ik})\} = X_{ik}\alpha_k + G_{ik}\beta_k$$

- Estimate β_k and its standard error, $(\widehat{\beta}_k, \widehat{\sigma}_k)$

$$\widehat{\beta}_k \sim N(\beta_k, \sigma_k^2)$$

Fixed Effect Model

- Assume $\beta_1 = \cdots = \beta_K$
- Effect size estimation

$$\widehat{\beta} = \frac{\sum_{k=1}^{K} w_k \widehat{\beta}_k}{\sum_{k=1}^{K} w_k}$$

$$\sqrt{n}(\widehat{\beta} - \beta) \to N\left(0, \frac{n \sum_k w_k^2 \sigma_k^2}{(\sum_k w_k)^2}\right)$$

- $w_k = \frac{1}{\mathrm{var}(\widehat{\beta}_k)}$ gives the minimal variance of $\widehat{\beta}$

Fixed Effect Model

- Assumes all studies in the analysis have the same effect
- Each study can be considered as a random sample drawn from the population with true parameter value β.
- There is between-study heterogeneity
  - Different definitions of phenotypes.
  - Effect size may be higher (or lower) in certain subgroups (e.g., age, sex).

Assess Heterogeneity

- Cochran's Q test

$$Q = \sum_{k=1}^{K} w_k(\widehat{\beta}_k - \widehat{\beta})^2$$

- Q should be large if there is heterogeneity.
- Under the null hypothesis of no heterogeneity

$$Q \sim \chi^2_{K-1}$$

- Under powered when there are fewer studies.

Assess Heterogeneity

- Measure the % of total variance explained by the between-study heterogeneity (Higgins and Thompson 2002)

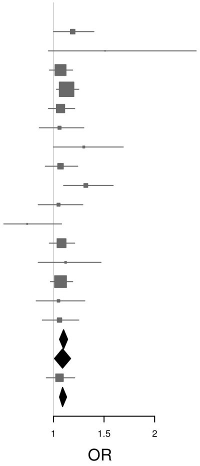
$$I^2 = \frac{Q - (K - 1)}{Q} \times 100\%$$

- $>$ 50% indicates large heterogeneity
- Intuitive interpretation, simple to calculate, and can be accompanied by an uncertainty interval

An Example

rs10911251

Study	**OR**	**95%CI**	**P**
ASTERISK	1.19	(1.00–1.40)	4.52e–02
COLO23	1.51	(0.95–2.41)	8.46e–02
CCFR	1.07	(0.96–1.19)	2.22e–01
DACHS	1.13	(1.03–1.25)	7.97e–03
DALS	1.07	(0.95–1.21)	2.65e–01
HPFS	1.06	(0.86–1.30)	5.73e–01
MEC	1.30	(1.00–1.69)	5.09e–02
NHS	1.07	(0.92–1.24)	3.84e–01
OFCCR	1.32	(1.10–1.59)	2.76e–03
PHS	1.05	(0.85–1.29)	6.58e–01
PMH	0.74	(0.51–1.08)	1.16e–01
PLCO	1.08	(0.96–1.21)	2.19e–01
VITAL	1.12	(0.85–1.47)	4.37e–01
WHI	1.07	(0.97–1.19)	1.83e–01
HPFS Ad	1.05	(0.83–1.31)	6.96e–01
NHS Ad	1.06	(0.89–1.25)	5.28e–01
GWAS	**1.10**	**(1.06–1.14)**	**1.34e–06**
Asian Follow–up	**1.09**	**(1.01–1.17)**	**3.20e–02**
Adenoma Follow–up	1.06	(0.93–1.21)	3.66e–01
GWAS+Follow–up	**1.09**	**(1.06–1.13)**	**9.45e–08**



Het pv=0.687

Meta- vs. Pooled-Analysis

- Meta analysis: Combining summary statistics (e.g., $\widehat{\beta}_k$) of studies
- Pooled analysis: Combining original or individual-level of all studies

Relative Efficiency

- For $k = 1, \cdots, K$ studies

$$g\{E(Y_{ki})\} = \alpha_k + \beta^T X_{ki}$$

- β is common to all K studies; α_k is specific to the kth study
- The maximum likelihood estimator (MLE) $\widehat{\beta}_k$ for the kth study by maximizing

$$L_k = \sup_{\alpha_k} \prod_{i=1}^{n_k} f(Y_{ki}, X_{ki}; \beta, \alpha_k)$$

Relative Efficiency

- $\widetilde{\beta}$ is MLE of β by maximizing

$$\begin{aligned} L &= \sup_{\{\alpha_1,\cdots,\alpha_K\}} \prod_{k=1}^{K} \prod_{i=1}^{n_k} f(Y_{ki}, X_{ki}; \beta, \alpha_k) \\ &= \prod_{k=1}^{K} \sup_{\alpha_k} \prod_{i=1}^{n_k} f(Y_{ki}, X_{ki}; \beta, \alpha_k) \\ &= \prod_{k=1}^{K} L_k \end{aligned}$$

Relative Efficiency

- Information

$$\mathrm{I}(\beta) = \frac{\partial^2}{\partial\beta^2}\log L = \sum_{k=1}^{K}\frac{\partial^2}{\partial\beta^2}\log L_k = \sum_{k=1}^{K}\mathrm{I}_k(\beta)$$

- Recall

$$\mathrm{var}(\widehat{\beta}) = \frac{1}{\sum_{k=1}^{K}\frac{1}{\mathrm{var}(\beta_k)}} = \frac{1}{\sum_{k=1}^{K}\mathrm{I}_k(\beta)}$$

- $\mathrm{var}(\widetilde{\beta}) = \frac{1}{I(\beta)}$
- Using summary statistics has the same asymptotic efficiency as using original data, if β is the only common parameter across studies.

Relative Efficiency

- Common nuisance parameters, say γ.

$$
\mathrm{I}_k = \begin{pmatrix} \mathrm{I}_{K\beta\beta} & \mathrm{I}_{K\beta\gamma} \\ \mathrm{I}_{K\partial\beta} & \mathrm{I}_{K\gamma\gamma} \end{pmatrix} \qquad \mathrm{I} = \begin{pmatrix} \mathrm{I}_{\beta\beta} & \mathrm{I}_{\beta\gamma} \\ \mathrm{I}_{\gamma\beta} & \mathrm{I}_{\gamma\gamma} \end{pmatrix}
$$

$$
\begin{aligned}
\mathrm{var}(\widehat{\beta}) &= \left\{ \sum_{k=1}^{K} (\mathrm{I}_{k\beta\beta} - \mathrm{I}_{k\beta\gamma}\mathrm{I}_{k\gamma\gamma}^{-1}\mathrm{I}_{k\gamma\beta}) \right\}^{-1} \\
&= (\mathrm{I}_{\beta\beta} - \sum_{k=1}^{K} \mathrm{I}_{k\beta\gamma}\mathrm{I}_{k\gamma\gamma}^{-1}\mathrm{I}_{k\gamma\beta})^{-1} \\
&\geq (\mathrm{I}_{\beta\beta} - \mathrm{I}_{\beta\gamma}\mathrm{I}_{\gamma\gamma}^{-1}\mathrm{I}_{\gamma\beta})^{-1} = \mathrm{var}(\widetilde{\beta})
\end{aligned}
$$

- Equality holds if and only of $\mathrm{var}(\beta_k)^{-1}\mathrm{cov}(\widehat{\beta}_k, \widehat{\gamma}_k)$ are the same among the K studies.

Random Effects Model

- Under the fixed effect model, $\beta_1 = \cdots = \beta_K = \beta$
- The random effects model

$$\beta_k = \beta + \xi_k \qquad (k = 1, \cdots, K)$$

- $\xi_k \sim N(0, \tau^2)$

Random Effects Model

- Estimation of τ^2
  - DerSimonian & Laird (1986) method-of-moments estimator

$$\widehat{\tau}^2 = \frac{Q-(K-1)}{\sum V_K^{-1} - \sum V_K^{-2} / \sum V_K^{-1}}$$

  - $V_k = \text{var}(\widehat{\beta}_k | \beta_k)$
- $\text{var}(\widehat{\beta}_k) = V_k + \tau^2$
- Estimate β with inverse variance estimator $\widehat{w}_k = \widehat{\text{var}}(\widehat{\beta}_k)$

$$\widehat{\beta} = \frac{\sum_{k=1}^{K} \widehat{w}_k \widehat{\beta}_k}{\sum_{k=1}^{K} \widehat{w}_k}$$

$$\text{SE}(\widehat{\beta}) = \frac{1}{\sqrt{\sum_k \widehat{w}_k}}$$

Random Effects Model

- Fixed effect model is more powerful, but ignores the heterogeneity between studies.
- Random effects model is probably more robust, but is under powered. The confidence intervals have poor coverage for small and moderate K.

Meta- vs Pooled-Analysis

- The maximum likelihood estimator $\widetilde{\beta}$ from pooled data can be obtained by maximizing

$$\prod_{k=1}^{K} \int f_k(Y_{ki}|X_{ki}, \beta + \xi_k) \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{\xi_k^2}{2\tau^2}\right) d\xi_k$$

- Challenging to establish the theoretical properties of $\widehat{\beta}$ and $\widetilde{\beta}$ under the random effects model because $n_k \gg K$.

Asymptotic Distributions (Zeng and Lin 2015)

- Assumptions:
    - For $k = 1, \cdots, K$, $n_k = \pi_k n$ for some constant π_k within a compact interval $\in (0, \infty)$.
    - $\tau^2 = \frac{1}{n}\sigma^2$, σ^2 is a constant. The between-study variability is comparable to the within-study variability.
- Asymptotic distribution of MLE $\widetilde{\beta}$

$$\sqrt{n}(\widetilde{\beta} - \beta_0) \to_d \left(\sum_{k=1}^{K} \frac{\pi_k}{v_k + \pi_k \mathcal{A}} \right) \sum_{k=1}^{K} \frac{\pi_k^{1/2}}{v_k + \pi_k \mathcal{A}} \mathcal{Z}_k$$

- $n\widetilde{\tau}^2 \to_d \widetilde{\mathcal{A}}; \quad n_K \widehat{V}_k \to v_k$
- $\mathcal{Z}_k$ are independently distributed $\sim N(0, v_k + \pi_k \sigma_0^2)$.
- $\mathcal{A}$ is a complicated form that involves $\{\mathcal{Z}_k\}$.
- It is a mixture of normal random variables with the mixing probabilities both being random and correlated with the normal random variables.

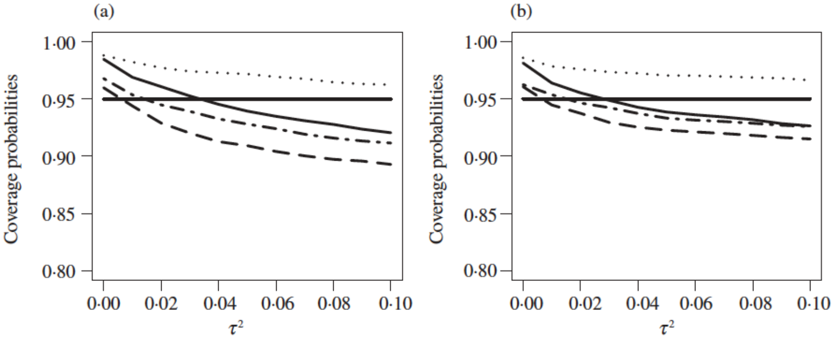
Asymptotic Distributions

- Asymptotic distribution of $\widehat{\beta}$

$$\sqrt{n}(\widehat{\beta} - \beta_0) \longrightarrow_d (\sum_{k=1}^{K} \frac{\pi_k}{v_k + \pi_k \widehat{\mathcal{A}}})^{-1} \sum_{k=1}^{K} \frac{\pi_k^{1/2} \mathcal{Z}_k}{v_k + \pi_k \widehat{\mathcal{A}}}$$

- $n\widehat{\tau}^2 \rightarrow_d \widehat{\mathcal{A}}$
- As $n, K \rightarrow \infty$, $Kn^{-1/2} \rightarrow 0$, $\quad \text{var}(\widetilde{\beta}) \geq \text{var}(\widehat{\beta})$, i.e., the weighted average estimator $\widehat{\beta}$ is at least as efficient as the MLE
  - When K = 100, 200, 400, the empirical relative efficiency is 1.037, 1.083, and 1.171.
- Perform statistical inference for $\widehat{\beta}$ and $\widetilde{\beta}$ based on asymptotic distributions using resampling techniques
- Or simpler, use the profile likelihood to construct 95% confidence intervals (Hardy and Thompson, 1996)

95% Coverage Probabilities



- Solid: Zeng & Lin; Dashed: DerSimonian-Laird; dotted Jackson-Bowden resampling method; dot-dash: Hardy-Thompson profile method. Left: K=10 studies; Right: K=20

Testing with random effects model

- Random effects (RE) model gives less significant *p* values than the fixed effects model when the variants show varying effect sizes between studies
- Ironic because RE is designed specifically for the case in which there is heterogeneity
- All associations identified RE are usually identified by the fixed effect model
- Causal variants showing high between-study heterogeneity might not be discovered by either method

Revisit the Traditional RE

- First step: estimate the effect size and CI by taking heterogeneity into account.
- Second step: normalize $\widehat{\beta}/\mathrm{SE}(\widehat{\beta})$ and translate it into *p*-value.
- Effectively, RE assumes heterogeneity under the null hypothesis, i.e.,

$$\widehat{\beta}_k \sim N(0, \sigma^2 + \tau^2).$$

- There should not be heterogeneity under the null because $\beta_1 = \cdots = \beta_K = 0$.

New RE

- New null hypothesis $H_0 : \beta = 0, \tau^2 = 0$.

$$\widehat{\beta}_k \sim N(0, \sigma^2)$$

- Model $\widehat{\beta} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_K)$.
    - $\widehat{\beta}$ follows a multivariate normal distribution

$$\widehat{\beta} \sim \mathrm{MVN}(\beta 1\ ,\ \Sigma)$$
$$\Sigma = V + \tau^2 \mathrm{I}\ , \qquad V = \mathrm{diag}(\,V_1, \cdots, V_K\,).$$

- The likelihood ratio test statistic.

$$S_{\text{new}} = -2 \log \frac{L_1}{L_0}$$
$$L_0 = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi V_k}} \exp\left(-\frac{\widehat{\beta}_k^2}{2V_k}\right)$$
$$L_1 = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi(V_k + \tau^2)}} \exp\left(-\frac{(\widehat{\beta}_k - \beta)^2}{2(V_k + \tau^2)}\right)$$

New RE

- Basically we test both fixed and random effects together

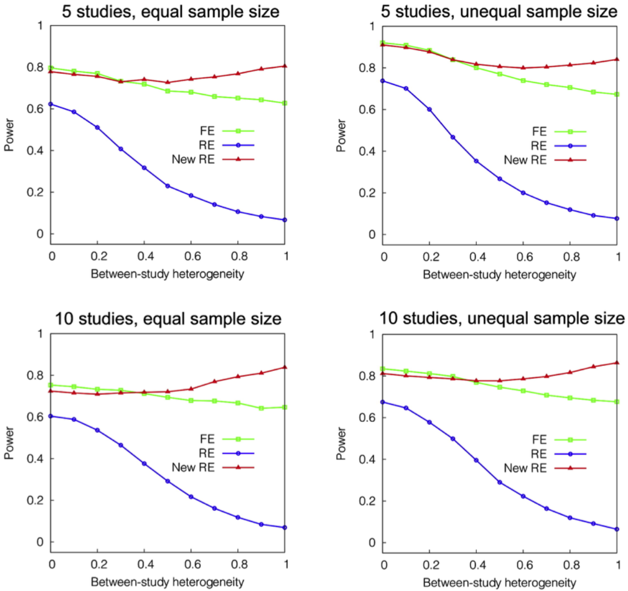
$$S_{\text{new}} = S_{\beta} + S_{\tau^2}$$

- S_{β} is equal to the Z-statistic based on the fixed effect model, and S_{τ^2} is the test statistic for testing $\tau^2 = 0$
- β is unconstrained, but $\tau^2 \geq 0$. Hence,

$$S_{\text{new}} \sim \frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_0.$$

- However, asymptotics cannot be applied because of small K. The asymptotic p-value is conservative because of the tail of asymptotic distribution is thicker than that of the true distribution at the tail
- Resampling approach should be used to calculate *p*-values

Power Comparison



5 studies, equal sample size

5 studies, unequal sample size

10 studies, equal sample size

10 studies, unequal sample size

Interpretation and Prioritization

- In the usual meta-analysis where one collects similar studies and expects the common effect, the results found by the fixed effects model should be the top priority.
- An association showing large heterogeneity requires careful investigation (e.g., different pathways, LD pattern, different study populations).
- Effect size estimate and CI is the same as those in the current RE, but may be inconsistent between wide CI and statistically significant results.

Comparison between meta- vs pooled-analysis

	Meta	Pooled
Logistic	Easy	Difficult
	Time-efficient	Time-consuming
	Cheap	Costly
	Bias	
Publication	Possible	Unlikely
Exposure assessment	May not be comparable	Comparable
Confounders	May not be consistent	Consistent
	Efficiency	
Relatively large data	Similar	Similar
Sparse data	May be unstable	Better
Complex analysis (e.g. machine learning)	Difficult	Easy
Long-term benefit	Always requires coordination	Better

Meta- vs pooled-analysis for sparse data

- $\widehat{\beta}$ is unstable if data are sparse in each study. However, if the interest is only on testing $H_0 : \beta = 0$, there are ways to combine the studies with only summary statistics.
- Recall the score test:

$$\left[\frac{\partial}{\partial\beta}\log L(\beta,\widehat{\alpha}_0)\right]\Big|_{\beta=0} \mathrm{I}(\beta=0|\widehat{\alpha}_0)^{-1} \left[\frac{\partial}{\partial\beta}\log L(\beta,\widehat{\alpha}_0)\right]\Big|_{\beta=0} \sim \chi^2_1$$

$$\mathrm{I}(\beta=0|\widehat{\alpha}_0) = \left\{\mathrm{I}_{\beta\beta} - \mathrm{I}_{\beta\alpha}\mathrm{I}_{\alpha\alpha}^{-1}\mathrm{I}_{\alpha\beta}\right\}\Big|_{\beta=0,\widehat{\alpha}_0}$$

- $\left[\frac{\partial}{\partial\beta}\log L(\beta,\widehat{\alpha}_0)\right]\Big|_{\beta=0} = \sum_{k=1}^{K}\left[\frac{\partial}{\partial\beta}\log L_k(\beta,\widehat{\alpha}_0)\right]\Big|_{\beta=0}$
- $\mathrm{I}_{\beta\beta}|_{\beta=0,\widehat{\alpha}_0} = \sum_{k=1}^{K}\mathrm{I}_{k\beta\beta}\big|_{\beta=0,\widehat{\alpha}_0}$. Similar for $\mathrm{I}_{\beta\alpha}$, $\mathrm{I}_{\alpha\alpha}$, and $\mathrm{I}_{\alpha\beta}$

Summary

- *P*-value based combination.
- Fixed vs random effects models.
- Meta vs. pooled- analysis.
- New random effects testing.

Recommended Reading

- DerSimonian R & Laird N (1986). Meta-analysis in clinical trials. *Contr Clin Trials* 7: 177-88.
- Han B and Eskin E (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* 88: 586–598.
- Hardy RJ & Thomson SG (1996) A likelihood approach to meta-analysis with random effects. *Stat in Med* 30: 619–29.
- Lin DY and Zeng D (2010). On the relative efficiency of using summary statistics vs. individual level data in meta-analysis. *Biometrika* 97: 321–32.
- Zeng D and Lin DY (2015). On random-effects meta-analysis. *Biometrika* 102: 281–294.