Inferring parts from case-control meta-analysis

Matti Pirinen

6.12.2021

Suppose that we have observed regression coefficient $\left(\widehat{\beta}_F, s_F\right)$ for **full** data analysis and a coefficient $\left(\widehat{\beta}_A, s_A\right)$ for a sub-analysis where cases and controls of A are included among cases and controls of F, respectively. Denote by B a (hypothetical) case-control study whose cases are those cases of F that are not in A and whose controls are among the controls of F. We have not observed results for B and we want to infer

- estimate $\widehat{\beta}_B$
- its standad error s_B and
- correlation $\widehat{r}_{AB}$ between estimators $\widehat{\beta}_A$ and $\widehat{\beta}_B$,

in order to compare statistically the effects $\widehat{\beta}_A$ and $\widehat{\beta}_B$.

Standard error s_B

Let's denote by $n_B^{(\text{eff})} = \frac{n_B^{(0)} n_B^{(1)}}{n_B}$ the effective sample size of study B, where $n_B^{(0)}$ and $n_B^{(1)}$ are the numbers of controls and cases, respectively, and $n_B = n_B^{(0)} + n_B^{(1)}$ is the total sample size. (Note that sometimes definition of effective sample size has an additional multiplier of 4.) In a case-control study, $\frac{1}{s_B^2} \approx C_B n_B^{(\text{eff})}$, where C_B depends on minor allele frequency and covariates. Given that MAF and covariates are similar in both A and B, we assume that $C_B \approx C_A$. Hence, we approximate

$$s_B \approx s_A \sqrt{\frac{n_A^{(\text{eff})}}{n_B^{(\text{eff})}}}.$$

When both A and B are meta-analyses over K individual cohorts, we first sum the effective sample sizes over the cohorts as, e.g.,
$n_B^{(\text{eff})} = \sum_{k=1}^K n_{B,k}^{(\text{eff})}$.
When we have available both s_F and s_A, we can use the maximum of the two candidate approximations as a conservative estimate:

$$s_B \approx \max\left\{ s_F \sqrt{\frac{n_F^{(\text{eff})}}{n_B^{(\text{eff})}}}, s_A \sqrt{\frac{n_A^{(\text{eff})}}{n_B^{(\text{eff})}}} \right\}.$$

Correlation r_{AB}

The formula by Bhattacharjee et al. 2012 in AJHG https://www.sciencedirect.com/science/article/pii/S0002929712001590 gives the correlation in z-scores (and equivalently for regression coefficients β) between the two case-control studies A and B based on their overlap in samples as

$$r_{AB} = \sqrt{n_A^{(\text{eff})} n_B^{(\text{eff})}} \left(\frac{n_{AB}^{(11)}}{n_A^{(1)} n_B^{(1)}} + \frac{n_{AB}^{(00)}}{n_A^{(0)} n_B^{(0)}} \right),$$

where $n_{AB}^{(11)}$ and $n_{AB}^{(00)}$ are, respectively, the number of shared cases and shared controls between the studies. (Here we have assumed that no case of one study is a control of the other study but Bhattacharjee et al. 2012 have also extended their formula for such more complex situations.)

Suppose that the study A and hypothetical study B are meta-analyses where we know their sample overlaps in each cohort. We first compute the correlation for each cohort based on the sample counts using the formula above and then weight the correlation estimates as derived next.

Suppose that there are K cohorts (some of which may have samples only for one of A or B). The meta-analysis estimate for A is

$$\widehat{\beta}_A = w_{A,1}\widehat{\beta}_{A,1} + \ldots + w_{A,K}\widehat{\beta}_{A,K}$$

where $w_{A,k} = s_{A,k}^{-2}/\phi_A$, $s_{A,k}$ is the standard error of estimator $\widehat{\beta}_{A,k}$ and $\phi_A = \sum_{k=1}^{K} s_{A,k}^{-2}$ is the precision (= inverse of the variance) of the combined estimator $\widehat{\beta}_A$. We assume that there is no covariance between different cohorts, i.e. $\mathrm{cov}\left(\widehat{\beta}_{A,k}, \widehat{\beta}_{B,k'}\right) = 0$ when $k \neq k'$. Then

$$\begin{aligned}
\mathrm{cov}\left(\widehat{\beta}_A, \widehat{\beta}_B\right) = \mathrm{cov}\left(\sum_{k=1}^{K} w_{A,k}\widehat{\beta}_{A,k}, \sum_{k'=1}^{K} w_{B,k'}\widehat{\beta}_{B,k'}\right) &= \sum_{k=1}^{K} w_{A,k}w_{B,k}\mathrm{cov}\left(\widehat{\beta}_{A,k}, \widehat{\beta}_{B,k}\right) = \sum_{k=1}^{K} w_{A,k}w_{B,k}s_{A,k}s_{B,k}r_{AB,k} \\
&= \sum_{k=1}^{K} \frac{s_{A,k}^{-1}}{\phi_A}\frac{s_{B,k}^{-1}}{\phi_B} r_{AB,k} \\
&= \frac{1}{\phi_A\phi_B}\sum_{k=1}^{K} s_{A,k}^{-1}s_{B,k}^{-1}r_{AB,k}.
\end{aligned}$$

Since variance of the inverse-variance weighted meta-analysis estimator $\widehat{\beta}_A$ is $1/\phi_A$, we have that the correlation

$$\mathrm{cor}\left(\widehat{\beta}_A, \widehat{\beta}_B\right) = \sqrt{\phi_A\phi_B}\mathrm{cov}\left(\widehat{\beta}_A, \widehat{\beta}_B\right) = \frac{1}{\sqrt{\phi_A\phi_B}}\sum_{k=1}^{K} s_{A,k}^{-1}s_{B,k}^{-1}r_{AB,k}.$$

By assuming, as above, that $s_{B,k} = s_{A,k}\sqrt{\frac{n_{A,k}^{(\mathrm{eff})}}{n_{B,k}^{(\mathrm{eff})}}}$ we have that $\phi_B = \sum_{k=1}^{K} s_{B,k}^{-2} = \sum_{k=1}^{K} s_{A,k}^{-2}\left(\frac{n_{B,k}^{(\mathrm{eff})}}{n_{A,k}^{(\mathrm{eff})}}\right)$ and we can estimate the correlation r_{AB} by using the values $s_{A,k}, n_{A,k}^{(\mathrm{eff})}$ and $n_{B,k}^{(\mathrm{eff})}$.

If we do not have access to $s_{A,k}$, we need to assume that the standard errors of all cohorts are proportional to effective sample size with the same constant, and we have

$$\mathrm{cor}\left(\widehat{\beta}_A, \widehat{\beta}_B\right) \approx \frac{\sum_{k=1}^{K}\sqrt{n_{A,k}^{(\mathrm{eff})}n_{B,k}^{(\mathrm{eff})}}r_{AB,k}}{\sqrt{n_A^{(\mathrm{eff})}n_B^{(\mathrm{eff})}}}.$$

Effect estimate $\widehat{\beta}_B$

To estimate $\widehat{\beta}_B$, we imagine that the full analysis F coefficients result from combining the analyses A and B that together contain all the cases and controls of F (with possibly considerable overlap in controls). In practice, we assume that F results from the most efficient linear combination of A and B given the SEs and the correlation between A and B. The combination is defined by weight $w = w_A \in (0, 1)$ of study A as

$$\beta'_F = w\beta_A + (1 - w)\beta_B.$$

The variance of β'_F is

$$\mathrm{Var}\left(\widehat{\beta'_F}\right) = w^2 s_A^2 + (1 - w)^2 s_B^2 + 2w(1 - w)s_A s_B r_{AB}.$$

We estimate w by the condition that Var $\left(\widehat{\beta'_F}\right)$ is minimized. The minimum is at

$$w = \frac{s_B^2 - s_A s_B r_{AB}}{s_A^2 + s_B^2 - 2 s_A s_B r_{AB}}.$$

With the value of w, we can estimate

$$\widehat{\beta}_B = \frac{\widehat{\beta}_F - w\widehat{\beta}_A}{1 - w}.$$

When $r_{AB} = 0$, this corresponds to the standard fixed effect meta-analysis estimator.

Application to COVID-19 HGI data

In COVID-19 hgi, there are separate analyses done for

- C2 GWAS: all infected as cases vs. population controls,
- B2 GWAS: hospitalized COVID patients as cases vs. population controls.

However, no analysis is done for "non-hospitalized infected" vs. population controls, which would be interesting to determine which associations are affecting the severity of the disease and which are affecting susceptibility of the infection. Hence, we have applied the above methods to approximate an effect size for a hypothetical "C2-B2" GWAS where cases are corona virus infected individuals who have not been hospitalized because of COVID and controls are population controls.

Since we have had to made many strong assumptions about the sample overlap between B2 and C2 analyses, we emphasize that such results may give only a crude approximation and that therefore we are not strongly relying on them in any inference. For example, in the subtype assignment model, we have not used the inferred C2-B2 effect sizes, but only B2 and C2 effect sizes that were observed.

However, we have used the weight w, that describes what would be the meta-analysis weight of GWAS B2 in GWAS C2 when we consider C2 as being combined from B2 and C2-B2, to define the expected effect size seen in C2 GWAS at such variants that affect severity of the disease but not susceptibility to infection.

We note that the resulting value $w = 0.200$ is practically the same as the ratio of effective sample sizes of B2 and C2 (0.208) and hence our results about subtype assignment would not change in practice had we used simply the latter value rather than the former in the Bayesian model comparison framework.