Bayesian Meta Analysis with Multiple Data Sources

Fan Bu, Aki Nishimura and Marc Suchard

September 2021

Introduction In many areas of health science research, it is common that multiple studies are conducted on multiple observational data sources to investigate similar research questions, and it is also of interest to combine and summarize results across multiple studies. More importantly, different studies may employ different designs to estimate the effect. In addition, there are concerns about biases due to systematic errors in observational data, and previous works have introduced the use of negative controls to detect such biases and adjust the effect estimates accordingly (Lipsitch et al., 2010; Arnold et al., 2016; Schuemie et al., 2018). Our goal is to combine multiple studies while adjusting for biases based on analyses of negative controls in order to produce a more robust estimate for a quantity of interest. We accomplish this through a Bayesian hierarchical modeling framework.

Related work We build on recent work of combining estimates obtained from multiple sources (Yao et al., 2021) and extend existing work that performs calibration on estimates obtained from a single data source (Mulgrave et al., 2020). We note that existing work of estimate combination and/or empirical calibration has heavily focused on adjusting available parameter point estimates, but in this work we wish to address a more general case where we directly combine information from the likelihood functions (or proper approximation of the likelihoods) of the parameter of interest across data sources.

The model We develop a Bayesian hierarchical model to obtain combined estimates from multiple sources while adjusting for the systematic errors that induce biases at different sources using negative control outcomes. We assume that the effects (or parameters) of interest at different data sources are the same or equivalent, even though the study designs adopted at those sources can be different. For example, researchers may choose the case-control design or self-control design depending on the characteristics of different data sources, but the quantity of interest should all be the (log) odds ratio or (log) relative risk of a certain outcome between two exposures.

Let $\theta \in \mathbb{R}$ denote the common quantity of interest - for simplicity, we will refer to this quantity as the "effect" to estimate. Suppose there are N data sources (indexed by i), and for source i, let the true effect be θ_i. Due to systematic errors, observational data at source i can only provide us with the likelihood function $f_i(\tilde{\theta}_i)$ with respect to a "biased" effect $\tilde{\theta}_i$. Denote the bias related to the estimation of this quantity by b_i (which is unknown), then we have $\tilde{\theta}_i = \theta_i + b_i$.

Our goal is to acquire a summary estimate θ while adjusting for the source-specific bias b_i (and thus effectively acquiring de-biased estimates for θ_i's). We assume the following hierarchical structure for the source-specific (true) effect θ_i's:

$$\theta_i \sim N(\theta, \gamma^2), \tag{1}$$

which also means

$$\tilde{\theta}_i \sim N(\theta + b_i, \gamma^2). \tag{2}$$

To adjust for the biases, M negative controls (indexed by j) are selected and used in analyses across the data sources. Let y_{ij} denote the estimated effect (i.e., estimated bias) for the jth negative control

on the ith data source. Suppose that y_{ij} follows a bias distribution P_i with density function $p_i(\cdot)$. We further assume that the bias b_i also follows the same distribution. Denote the parameters of this distribution by $\boldsymbol{\mu}_i$. For example, if we believe that the bias distribution is normal, then $\boldsymbol{\mu}_i = (\beta_i, \tau_i^2)$ where β_i is the mean bias and τ_i^2 is the variance.

Inference The key parameters of interest include θ (the summary quantity) and also the θ_i's (the source-specific true effect). Additionally, the quantities $b_i, \gamma^2, \boldsymbol{\mu}$ are also unknown. Next we describe the Bayesian inference scheme for estimating the model parameters. Let π_0 and π_i denote the priors we adopt for (θ, γ^2) and $\boldsymbol{\mu}_i$, respectively. For instance, we can use the normal-inverse-Gamma prior for π_0 and do the same for π_i if we assume a normal bias distribution.

We may consider a Markov chain Monte Carlo inference scheme or simply use Hamiltonian Monte Carlo (e.g., through `stan`). The MCMC algorithm is sketched as follows:

In each iteration, do:

1. sample b_i from its full conditional $\propto p_i(b_i)\text{pnorm}(\tilde{\theta}_i; \theta + b_i, \gamma^2)$;
2. sample $\tilde{\theta}_i$ from its full conditional $\propto f_i(\tilde{\theta}_i)\text{pnorm}(\tilde{\theta}_i; \theta + b_i, \gamma^2)$;
3. sample θ, γ^2 for their full conditional $\propto \pi_0(\theta, \gamma^2) \prod_{i=1}^{N} \text{pnorm}(\tilde{\theta}_i - b_i; \theta, \gamma^2)$;
4. sample $\boldsymbol{\mu}$ from full conditional $\propto \pi_i(\boldsymbol{\mu}_i) p_i(b_i; \boldsymbol{\mu}_i) \prod_{j=1}^{M} p_i(y_{ij}; \boldsymbol{\mu}_i)$.

Here "pnorm" represents the normal density function. We note that if a normal model for the biases is assumed (i.e., P_i is normal), then Steps 1, 3 and 4 become trivial updates of hierarchical normal models if we adopt semi-conjugate normal-inverse-Gamma priors. The only potential bottleneck is in Step 2, where we may need to run a Metropolis-Hastings step.

Furthermore, it is possible to directly utilize an estimated (empirical) bias distribution P_i if we so wish. In this case, we can keep density function p_i fixed and only run the first three steps in the above algorithm.

Discussion We have assumed a normal distribution in (1) for simplicity and also based on previous findings that normal distributions (or approximations) seem to work well in practice (Schuemie et al., 2018; Mulgrave et al., 2020). However, for cases where the normality assumption doesn't hold (e.g., when effects from different sources have a larger dispersion), non-normal models like normal mixtures or t distributions could be considered. Further, we may consider different transformations for other types of quantities of interest - for example, if the parameter to estimate is a proportion, then we can adopt a logit-normal model or a Beta distribution instead of the log-normal model assumed here.

References

Arnold, B. F., A. Ercumen, J. Benjamin-Chung, and J. M. Colford Jr (2016). Brief report: negative controls to detect selection bias and measurement bias in epidemiologic studies. *Epidemiology (Cambridge, Mass.) 27*(5), 637.

Lipsitch, M., E. T. Tchetgen, and T. Cohen (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.) 21*(3), 383.

Mulgrave, J. J., D. Madigan, and G. Hripcsak (2020). Bayesian posterior interval calibration to improve the interpretability of observational studies. *arXiv preprint arXiv:2003.06002*.

Schuemie, M. J., G. Hripcsak, P. B. Ryan, D. Madigan, and M. A. Suchard (2018). Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences 115*(11), 2571–2577.

Yao, Y., R. T. Ogden, C. Zeng, and Q. Chen (2021). Bivariate hierarchical bayesian model for combining summary measures and their uncertainties from multiple sources. *arXiv preprint arXiv:2109.07560*.