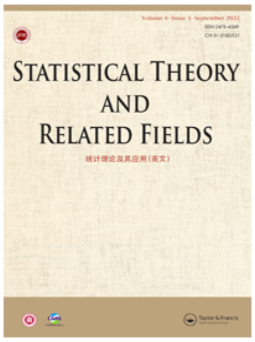

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

A discussion on “A selective review of statistical methods using calibration information from similar studies” by Qin, Liu and Li

Peisong Han

To cite this article: Peisong Han (2022) A discussion on “A selective review of statistical methods using calibration information from similar studies” by Qin, Liu and Li, Statistical Theory and Related Fields, 6:3, 193-195, DOI: 10.1080/24754269.2022.2084929

To link to this article: https://doi.org/10.1080/24754269.2022.2084929

Published online: 10 Jun 2022.

STATISTICAL THEORY AND RELATED FIELDS
2022, VOL. 6, NO. 3, 193–195
https://doi.org/10.1080/24754269.2022.2084929

SHORT COMMUNICATION

OPEN ACCESS

A discussion on “A selective review of statistical methods using calibration information from similar studies” by Qin, Liu and Li

Peisong Han

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA

ARTICLE HISTORY Received 20 April 2022; Revised 30 May 2022; Accepted 2 May 2022

We congratulate Qin, Liu and Li (QLL) on a thoughtful and much needed review of many interesting methods for combining information from similar studies. We appreciate being given the opportunity to make a discussion. QLL cover a variety of different settings and methods. Based on that, we will provide a brief review on some additional relevant literature with a focus on methods that deal with population heterogeneity, since in practice it is most likely that different studies sample from different populations and whether information should be combined depends on how similar those populations are, among many other considerations. To keep the discussion focussed, we will follow the setting in Section 5 of QLL, although most of these methods can be more broadly applied.

We adopt the notation in Section 5.1 of QLL with some variations. Let $(Y_i, \boldsymbol{X}_i^{\mathrm{T}}, \boldsymbol{Z}_i^{\mathrm{T}})^{\mathrm{T}}$, $i = 1, \ldots, n$, denote the individual data based on a random sample from the internal study, where Y is the response and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are vectors of covariates. The model of interest is $f(Y \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ for $f(Y \mid \boldsymbol{X}, \boldsymbol{Z})$ with parameter $\boldsymbol{\beta}$. The external study fitted a (possibly misspecified) model $f(Y \mid \boldsymbol{X}; \boldsymbol{\theta})$ for $f^*(Y \mid \boldsymbol{X})$ with covariates $\boldsymbol{X}$ alone and parameter $\boldsymbol{\theta}$. Throughout this discussion we use a $*$-superscript to denote distributions/expectations/quantities associated with the external study population. The external model fitting information is summarized by

$$E^*\{\boldsymbol{h}(Y, \boldsymbol{X}; \boldsymbol{\theta}^*)\} = \boldsymbol{0}, \tag{1}$$

where $\boldsymbol{h}(Y, \boldsymbol{X}; \boldsymbol{\theta})$ is the score function for $f(Y \mid \boldsymbol{X}; \boldsymbol{\theta})$ and $\boldsymbol{\theta}^*$ is the solution to the score equation based on the external study sample. Individual data from the external sample are not available. We assume the external sample size is very large so that the uncertainty in $\boldsymbol{\theta}^*$ is negligible compared to the internal study, i.e. Case I in QLL.

QLL in Section 5 give an excellent review of some methods and their comparisons in terms of asymptotic efficiency when the internal and external study populations are the same, based on both the empirical likelihood (EL) formulation (Owen, 1988; Qin & Lawless, 1994) and the constrained maximum likelihood (CML) formulation (Chatterjee et al., 2016; Qin, 2000). These two formulations are closely connected (Han & Lawless, 2016). For ease of discussion, we provide the CML formulation here, which is already covered in QLL. When $f(Y, \boldsymbol{X}, \boldsymbol{Z}) = f^*(Y, \boldsymbol{X}, \boldsymbol{Z})$, (1) can be transformed into

$$E\{\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})\} = \boldsymbol{0}, \tag{2}$$

where

$$\begin{aligned}\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta}) &= E\{\boldsymbol{h}(Y, \boldsymbol{X}; \boldsymbol{\theta}^*) \mid \boldsymbol{X}, \boldsymbol{Z}\} \\ &= \int \boldsymbol{h}(Y, \boldsymbol{X}; \boldsymbol{\theta}^*) f(Y \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})\, \mathrm{d}Y. \end{aligned} \tag{3}$$

The CML estimator $\widehat{\boldsymbol{\beta}}_{cml}$ is defined through

$$\begin{aligned} &\max_{\boldsymbol{\beta}, p_1, \ldots, p_n} \quad \prod_{i=1}^{n} f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) p_i \\ &\text{subject to} \quad p_i \geq 0, \ \sum_{i=1}^{n} p_i = 1, \\ &\qquad\qquad \sum_{i=1}^{n} p_i \boldsymbol{\psi}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) = \boldsymbol{0}, \end{aligned} \tag{4}$$

where $p_i \equiv \mathrm{d}F(\boldsymbol{X}_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$.

When $f(Y \mid \boldsymbol{X}, \boldsymbol{Z}) = f^*(Y \mid \boldsymbol{X}, \boldsymbol{Z})$ but $f(\boldsymbol{X}, \boldsymbol{Z}) \neq f^*(\boldsymbol{X}, \boldsymbol{Z})$, applying the transformation (3) to (1) will lead to

$$E^*\{\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})\} = \boldsymbol{0}, \tag{5}$$

where the expectation $E^*(\cdot)$ is under $f^*(\boldsymbol{X}, \boldsymbol{Z})$ in contrast to the $E(\cdot)$ in (2) that is under $f(\boldsymbol{X}, \boldsymbol{Z})$. In this case, although (2) no longer holds, there are ways to make use of the external study information summarized by (5). One way is to collect a small supplementary sample $(\boldsymbol{X}_j^{*\mathrm{T}}, \boldsymbol{Z}_j^{*\mathrm{T}})^{\mathrm{T}}$, $j = 1, \ldots, n^*$, from the external

CONTACT Peisong Han peisong@umich.edu

study population (Chatterjee et al., 2016; Han & Lawless, 2019), and define the CML estimator through

$$\max_{\boldsymbol{\beta},p_1^*,\ldots,p_{n^*}^*} \quad \prod_{i=1}^{n} f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) \prod_{j=1}^{n^*} p_j^*$$

$$\text{subject to} \quad p_j^* \geq 0, \ \sum_{j=1}^{n^*} p_j^* = 1,$$

$$\sum_{j=1}^{n^*} p_j^* \boldsymbol{\psi}(\boldsymbol{X}_j^*, \boldsymbol{Z}_j^*; \boldsymbol{\beta}) = \boldsymbol{0},$$

where $p_j^* \equiv \mathrm{d}F^*(\boldsymbol{X}_j^*, \boldsymbol{Z}_j^*), j = 1, \ldots, n^*$. Another way is to specify a density ratio model of the form $f^*(\boldsymbol{X}, \boldsymbol{Z}) = \exp(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\alpha})f(\boldsymbol{X}, \boldsymbol{Z})$, where $\boldsymbol{W}$ is a subset of $(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Z}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\alpha}$ is newly introduced parameter, to link the external distribution $f^*(\boldsymbol{X}, \boldsymbol{Z})$ to the internal distribution $f(\boldsymbol{X}, \boldsymbol{Z})$ (Sheng et al., 2021). This approach models the heterogeneity between $f^*(\boldsymbol{X}, \boldsymbol{Z})$ and $f(\boldsymbol{X}, \boldsymbol{Z})$ by $\exp(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\alpha})$, and is in the same spirit that an exponential tilting model is specified to link the case distribution to the control distribution as in Section 7.1 of QLL. With the density ratio model, the CML estimator can be defined through

$$\max_{\boldsymbol{\beta},\boldsymbol{\alpha},p_1,\ldots,p_n} \quad \prod_{i=1}^{n} f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) p_i$$

$$\text{subject to} \quad p_i \geq 0, \ \sum_{i=1}^{n} p_i = 1,$$

$$\sum_{i=1}^{n} p_i \{\exp(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{\alpha}) - 1\} = \boldsymbol{0},$$

$$\sum_{i=1}^{n} p_i \boldsymbol{\psi}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) \exp(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{\alpha}) = \boldsymbol{0},$$

where $p_i \equiv \mathrm{d}F(\boldsymbol{X}_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$. Here the last two constraints are based on the fact that $f^*(\boldsymbol{X}, \boldsymbol{Z}) = \exp(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\alpha})f(\boldsymbol{X}, \boldsymbol{Z})$ is a density and (5), respectively. Note that the dimension of $\boldsymbol{\alpha}$ needs to be no larger than the dimension of $\boldsymbol{\psi}$ plus one for $\boldsymbol{\alpha}$ to be identifiable.

When $f(Y \mid \boldsymbol{X}, \boldsymbol{Z}) \neq f^*(Y \mid \boldsymbol{X}, \boldsymbol{Z})$, applying transformation (3) to (1) leads to neither (2) nor (5) in general, regardless of if $f(\boldsymbol{X}, \boldsymbol{Z})$ is the same as $f^*(\boldsymbol{X}, \boldsymbol{Z})$. In this case, the aforementioned CML estimators are biased. However, since the external study sample size is large, the reduction in variance by making use of external summary information may still benefit the internal study parameter estimation from a mean squared error perspective. Based on consideration of such a bias-variance trade-off, Estes et al. (2018) proposed an empirical Bayes shrinkage estimator of the form

$$\widehat{\boldsymbol{V}}_0(\widehat{\boldsymbol{V}}_{mle} + \widehat{\boldsymbol{V}}_0)^{-1}\widehat{\boldsymbol{\beta}}_{mle} + \widehat{\boldsymbol{V}}_{mle}(\widehat{\boldsymbol{V}}_{mle} + \widehat{\boldsymbol{V}}_0)^{-1}\widehat{\boldsymbol{\beta}}_{cml}$$

that is a weighted average of $\widehat{\boldsymbol{\beta}}_{mle}$, the MLE using the internal study data alone, and $\widehat{\boldsymbol{\beta}}_{cml}$ defined through (4). Here $\widehat{\boldsymbol{V}}_{mle}$ and $\widehat{\boldsymbol{V}}_0$ are estimated variance matrices for $\widehat{\boldsymbol{\beta}}_{mle}$ and for the prior normal distribution on $\boldsymbol{\beta}$, respectively. This method shrinks the final estimate towards $\widehat{\boldsymbol{\beta}}_{mle}$ in the presence of population heterogeneity and towards $\widehat{\boldsymbol{\beta}}_{cml}$ otherwise. Gu et al. (2021) extended the idea in Estes et al. (2018) to the case of multiple external studies, and propose an estimator that is a weighted average of the empirical Bayes estimators resulted from using each external study separately.

To deal with arbitrary population heterogeneity when information is available from multiple external studies but without causing estimation bias, Zhai and Han (2022) developed an estimation procedure that simultaneously selects the studies that give (2) and incorporates the corresponding information into internal model fitting. Their method also applies under the current setting of only one external study. When (2) does not hold because of an unknown form of heterogeneity, some components of $E\{\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})\}$ may still be zero even if $E\{\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})\}$ is not a zero vector, and these components still contain useful information to improve internal model estimation efficiency. This observation makes sense because the association between the same response and certain covariates may not differ much across populations with certain specific heterogeneity. Some general examples on this observation are given in Zhai and Han (2022). Let $\boldsymbol{\gamma} \equiv E\{\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})\}$, then the components of $\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ that correspond to zero-components of $\boldsymbol{\gamma}$ should be selected to compute $\widehat{\boldsymbol{\beta}}_{cml}$. To shrink the estimate of the zero-components of $\boldsymbol{\gamma}$ to exactly zero for information integration, under the current setting, the estimator in Zhai and Han (2022) is defined through

$$\max_{\boldsymbol{\beta},\boldsymbol{\gamma}} \quad \left[\max_{p_1,\ldots,p_n} \log \left\{ \prod_{i=1}^{n} f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) p_i \right\} - n \sum_{k=1}^{\dim(\boldsymbol{\gamma})} \lambda_n \frac{|\gamma_k|}{|\widetilde{\gamma}_k|^w} \right]$$

$$\text{subject to} \quad p_i \geq 0, \ \sum_{i=1}^{n} p_i = 1,$$

$$\sum_{i=1}^{n} p_i \{\boldsymbol{\psi}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\beta}) - \boldsymbol{\gamma}\} = \boldsymbol{0},$$

with an adaptive Lasso penalty (Zou, 2006) on $\boldsymbol{\gamma}$ with tuning parameter $\lambda_n > 0$, where $\widetilde{\gamma}_k$ is the kth component of $\widetilde{\boldsymbol{\gamma}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{X}_i, \boldsymbol{Z}_i; \widehat{\boldsymbol{\beta}}_{mle})$ and $w > 0$ is some user-specified positive number typically taken to be 1 or 2. Similar idea has also been considered for integrating external summary information into survival analysis with different penalty (Chen et al., 2021).

All of the aforementioned methods make use of the external summary information through transformation (3). Taylor et al. (2022) took a different approach when both $f(Y \mid \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$ and $f(Y \mid \boldsymbol{X}; \boldsymbol{\theta})$ are generalized linear models (GLM), namely

$$g\{\mathbb{E}(Y\,|\,\boldsymbol{X},\boldsymbol{Z})\} = \beta_0 + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_X + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_Z \qquad (6)$$

and $l\{\mathbb{E}(Y\,|\,\boldsymbol{X})\} = \theta_0 + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}_X$, with possibly different link functions $g(\cdot)$ and $l(\cdot)$ (note that the second GLM may be misspecified). Here the notation $\mathbb{E}(\cdot)$ for expectation is generic to simply present the form of the model. For ease of discussion, assume covariates $\boldsymbol{Z}$ have been orthogonalized to $\boldsymbol{X}$, which can be done by taking $\boldsymbol{Z}$ to be the vector of residuals of the least squares regression of each covariate in $\boldsymbol{Z}$ on $\boldsymbol{X}$ using the internal data. Taylor et al. (2022) showed that $\boldsymbol{\beta}_X \approx c\boldsymbol{\theta}_X$ for some unknown constant c when both GLMs are fitted to the same infinitely large sample and when $\boldsymbol{\beta}_X$, $\boldsymbol{\beta}_Z$ and $\boldsymbol{\theta}_X$ are all close to zero (here in this sentence, with some abuse of notation, $\boldsymbol{\beta}_X$, $\boldsymbol{\beta}_Z$ and $\boldsymbol{\theta}_X$ are the values after fitting both GLMs to the same infinitely large sample). See also Neuhaus and Jewell (1993). Based on this result, when $f(Y\,|\,\boldsymbol{X})$ and $f^*(Y\,|\,\boldsymbol{X})$ are similar in the sense that the relative effects of $\boldsymbol{X}$ on Y (but not necessarily the absolute magnitudes) are the same between the two populations, the $\boldsymbol{\theta}^*_X$ produced by the external study can still be used to improve the internal estimation efficiency. With $\boldsymbol{Z}$ orthogonalized to $\boldsymbol{X}$, instead of fitting (6), Taylor et al. (2022) proposed to fit

$$g\{\mathbb{E}(Y\,|\,\boldsymbol{X},\boldsymbol{Z})\} = \beta_0 + \alpha(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}^*_X) + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_Z$$

with coefficients $(\beta_0, \alpha, \boldsymbol{\beta}_Z^{\mathrm{T}})^{\mathrm{T}}$ to the internal study data, which is equivalent to letting $\boldsymbol{\beta}_X = \alpha\boldsymbol{\theta}^*_X$.

In the presence of different study populations, a crucial question to ask before combining information is which population is of the primary interest. Most of the methods reviewed in this discussion, if not all, explicitly or implicitly assume that the internal study population is of the primary interest and the external summary information is used for efficiency improvement without causing (too much) bias. This is reasonable, for example, when the internal study has a clear target population and is based on a careful design with a well controlled sampling. In practice, with data usually collected based on convenience sampling, the internal study sample may not be representative of the target population, or there may even be ambiguity about the target population itself. Therefore, some cautions are always needed when applying those methods to combine information in the presence of population heterogeneity, and more methodological developments are definitely needed.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Chatterjee, N., Chen, Y. H., Maas, P., & Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, *111*(513), 107–117. https://doi.org/10.1080/01621459.2015.1123157

Chen, Z., Ning, J., Shen, Y., & Qin, J. (2021). Combining primary cohort data with external aggregate information without assuming comparability. *Biometrics*, *77*(3), 1024–1036. https://doi.org/10.1111/biom.v77.3

Estes, J. P., Mukherjee, B., & Taylor, J. M. G. (2018). Empirical bayes estimation and prediction using summary-level information from external big data sources adjusting for violations of transportability. *Statistics in Biosciences*, *10*(3), 568–586. https://doi.org/10.1007/s12561-018-9217-4

Gu, T., Taylor, J. M. G., & Mukherjee, B. (2021). A meta-inference framework to integrate multiple external models into a current study. *Biostatistics*, kxab017. https://doi.org/10.1093/biostatistics/kxab017

Han, P., & Lawless, J. F. (2016). Discussion of "Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources". *Journal of the American Statistical Association*, *111*(513), 118–121. https://doi.org/10.1080/01621459.2016.1149399

Han, P., & Lawless, J. F. (2019). Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statistica Sinica*, *29*, 1321–1342.

Neuhaus, J. M., & Jewell, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, *80*(4), 807–815. https://doi.org/10.1093/biomet/80.4.807

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, *75*(2), 237–249. https://doi.org/10.1093/biomet/75.2.237

Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika*, *87*(2), 484–490. https://doi.org/10.1093/biomet/87.2.484

Qin, J., & Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, *22*(1), 300–325. https://doi.org/10.1214/aos/1176325370

Sheng, Y., Sun, Y., Huang, C.-Y, & Kim, M.-O. (2021). Synthesizing external aggregated information in the presence of population heterogeneity: A penalized empirical likelihood approach. *Biometrics*. https://doi.org/10.1111/biom.13429

Taylor, J. M. G., Choi, K., & Han, P. (2022). Data integration – exploiting ratios of parameter estimates from a reduced external model. *Biometrika*. https://doi.org/10.1093/biomet/asac022

Zhai, Y., & Han, P. (2022). Data integration with oracle use of external information from heterogeneous populations. *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2022.2050248

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735