Developing prediction models when there are systematically missing predictors in an individual patient data meta-analysis

Michael Seo[1,2]

August 10, 2021

[1]Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

[2]Graduate School for Health Sciences, University of Bern, Bern, Switzerland

Introduction

- Meta-analysis of individual patient data (IPD) is widely used to synthesize patient-level data from multiple studies, when aiming to estimate relative treatment effects or when developing clinical prediction models.
- One practical problem is when we have predictors that are only partially reported in studies. We refer to this situation as 'sporadically missing' data.
- Issue that we will focus on is when predictors are **'systematically missing'**, which is when predictors are missing entirely in some studies.

Introduction (2)

- One method for handling both types of missing data is multiple imputation (MI).
- When meta-analyzing IPD from multiple studies, MI needs to take into account the multilevel structure.
- When imputing systematically missing predictors, it is essential to allow for potential heterogeneity in different studies.
- We explore different methods of addressing the systematically missing predictors problem, when the aim is to **build a prediction model using data from multiple studies**.

One-stage meta-analytical prediction model

- In case all studies report on all predictors we can fit a one-stage meta-analytical prediction model.
- Assume continuous outcome $y_{ij} \sim Normal(\mu_{ij}, \sigma^2)$ where μ_{ij} denotes the expected outcome of the patient i randomized in study j.
- The simplest model consists only of linear terms:

$$\mu_{ij} = a_j + \boldsymbol{b}\boldsymbol{x}_{\boldsymbol{ij}}$$

$$a_j \sim N(\alpha, \tau_\alpha^2)$$

where a_j is the study-specific intercept, assumed to be normally distributed across studies with a mean α and variance τ_α^2 and $\boldsymbol{b}$ denotes the regression coefficients of the main effects of the covariates $\boldsymbol{x}_{\boldsymbol{ij}}$.

Simple one-stage meta-analytical prediction model (2)

- Often when meta-analyzing IPD of randomized controlled trials (RCTs), we include treatment effect term and/or treatment-covariate interactions.
- With these included, linear predictor is modelled as follows:

$$\mu_{ij} = a_j + \boldsymbol{b}x_{ij} + \boldsymbol{c}x_{ij}\, t_{ij} + d_j t_{ij}$$

$$a_j \sim N(\alpha, \tau_\alpha^2),\ d_j \sim N(\delta, \tau_\delta^2)$$

where $\boldsymbol{c}$ denotes the coefficients for treatment-covariate interactions, d_j is the study-specific treatment effect, assumed to be normally distributed across studies, with average δ and variance τ_δ^2, t_{ij} is the indicator for treatments.

Simple one-stage meta-analytical prediction model (3)

- After we fit the model, we can predict the outcome for a new patient with covariates $\boldsymbol{x}^{\boldsymbol{new}}$ using the estimated values for the parameters of the model. We simply use the average mean for the random effects.
- In the case of studies without treatment, we would have

$$y_{pred}(\boldsymbol{x}^{\boldsymbol{new}}, t) = \hat{\alpha} + \hat{\boldsymbol{\beta}}\boldsymbol{x}^{\boldsymbol{new}}$$

- In IPD of RCTs,

$$y_{pred}(\boldsymbol{x}^{\boldsymbol{new}}, t) = \hat{\alpha} + \hat{\boldsymbol{\beta}}\boldsymbol{x}^{\boldsymbol{new}} + \hat{\boldsymbol{\gamma}}\boldsymbol{x}^{\boldsymbol{new}} t + \hat{\delta} t$$

Naive method

- We consider different methods for addressing systematically missing predictors. In this first approach, we need all studies to report the same set of covariates.
- When this is not the case, we drop from the model development predictors that are not reported in one or more studies.
- Obviously, this approach will be wasteful, especially when the ignored predictors are important. Nonetheless, this is often done among researchers when analyzing IPD.
- After removing predictors, we fit and make predictions with the simple one-stage meta-analytical model.

Imputation method

- Instead of dropping covariates, we can use multiple imputation methods to impute systematically missing predictors.
- MI methods developed for sporadically missing data cannot be used to impute systematically missing data since variance-covariance matrix within cluster cannot be updated in the algorithm.
- There are three recently proposed methods for imputing systematically missing predictors: Jolani et al., Quartagno and Carpenter, and Resche-Rigon and White.
- All three approaches are viable options and the paper by Audigier et al. compares these methods extensively.

Imputation method (2)

- Under adequate sample size per cluster and few observed clusters, Resche-Rigon and White approach is recommended.
- This approach is a fully conditional specification imputation model based on a two-stage estimator.
- For each cluster without systematically missing data, the method estimates regression coefficients, variance-covariance matrix of the regression coefficients, and within-cluster variance of the outcome.
- Then, the method performs a multivariate meta-analysis using these cluster specific estimates to calculate variance-covariance matrix of the random effects.
- Ultimately, using random draws from the estimated distribution of these parameters, the model imputes a value for the systematically missing predictors.

Imputation method (3)

- Once systematically missing variables are multiply imputed, we make predictions using the imputated datasets and pool them using the Rubin's rules.
- There is a recent R package (developed by Audigier et al.) called *micemd* which implements these imputation methods.
- This is an extension of the *mice* package. We can use method = *2l.2stage.norm*, *2l.2stage.bin*, *2l.2stage.pois*, depending on the type of predictors imputed.

Separate prediction method

- Fit a different model for each study, i.e. using only the predictors reported in the study.
- This circumvents the problem of having studies not measuring certain predictors, without having to impute them.
- More specifically, we fit the model for each study separately:

$$\mu_i = a + \boldsymbol{bx_i}$$

$$\mu_i = a + \boldsymbol{bx_i} + \boldsymbol{cx_i}\, t_i + dt_i$$

Note that we no longer write index j for study and do not assume random effects for study intercept or treatment effect.

Separate prediction method (2)

- The set of predictors we use for each study will be different, as some studies do not report some of the predictors. This means that in principle we may end up fitting a different model in each study.
- After developing the models, we can use these N_S (total number of studies) models to predict the outcome for new patients.
- We combine these predictions into a single, final prediction.
- We can take a simple average or use a weighted average using precision of the prediction estimates.

Methods for assessing the model performance

- We use mean squared error, mean absolute error, and R-squared to evaluate a model's performance.
- We use leave-one-study-out cross validation method.
- In this procedure, one study is left out of the data and the rest of the studies are used to develop the model. Then, the fitted model is used to make predictions and measure performance in the patients of the left-out study.

Illustrative example

- We use a dataset of 12 trials in psychotherapies for depression, comparing treatment as usual (TAU) versus internet delivered cognitive behavioral therapy. The outcome is Patient Health Questionnaire-9 scores (PHQ-9).
- There are two continuous predictors (baseline PHQ-9 scores and age), five binary predictors (gender, relationship status, comorbid anxiety, medication, and alcohol use), and one count predictor (number of previous episodes).

Overview of the systematically missing data

***Table 1**: Overview of the systematically missing data for each study in the illustrative example.*

Study	Baseline	Gender	Age	Relationship status	Comorbid anxiety	Number of previous episodes	Medication	Alcohol
De Graaf 2009	✓	✓	✓	✓	✓	✓	✗	✓
Farrer 2011	✓	✓	✓	✓	✓	✓	✗	✓
Geraedts 2014	✓	✓	✓	✓	✓	✗	✗	✗
Gilbody 2015	✓	✓	✗	✓	✗	✗	✗	✗
Johansson 2012	✓	✓	✓	✓	✓	✗	✓	✗
Kivi 2014	✓	✓	✓	✓	✓	✓	✓	✓
Klein 2016 high	✓	✓	✓	✓	✓	✓	✓	✓
Klein 2016 low	✓	✓	✓	✓	✓	✓	✓	✓
Meyer 2015	✓	✓	✓	✓	✓	✓	✓	✗
Philips 2014	✓	✓	✓	✓	✗	✓	✓	✗
Montero-Marin 2016	✓	✓	✓	✓	✓	✗	✗	✗
Rosso 2016	✓	✓	✓	✗	✓	✓	✓	✗

✗ indicate systematically missing; ✓ indicate not systematically missing

Results

Table 2 *Performance metric result for the real dataset. The method with lowest MSE/MAE and highest R-squared is bolded. MSE: mean squared error; MAE: mean absolute error.*

Performance measure	Naïve	Imputation	Separate prediction
MSE	26.7212	**26.7131**	26.8937
MAE	**4.1266**	4.1424	4.1482
R-squared	0.1671	**0.1673**	0.1617

- Naive analysis only include baseline PHQ-9 score and gender.
- However, from regression estimates, baseline score is the most important covariate that explains most of the variance. This may explain why naive method performs well comparatively.

Simulation study

- We explored different configurations regarding the number of studies, the number of predictors, the number of systematically missing studies, the existence of non-linear predictor effects, the noise to signal ratio, the existence of treatment-covariate interactions, and extent of between-study variance of regression coefficients.
- We found that the **imputation method** worked best on average.
- However, under certain non-linear relationship in the outcome and predictors and when there are not enough observed studies to estimate between-study variance of regression coefficients correctly, **separate prediction method** sometimes had better performance over naive method or imputation method.

Conclusion

- We explored different methods for making predictions when there are systematically missing predictors in an individual patient data meta-analysis.
- Depending on the clinical settings, the added benefit of utilizing more advanced prediction method can be substantial and naive method should be avoided.
- The R codes used for fitting all models are available at: https://github.com/MikeJSeo/phd/tree/master/missing

References

Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., van Buuren, S., and Resche-Rigon, M. (2018). Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Statistical Science*, 33(2):160 – 183.

Debray, T. P. A., Moons, K. G. M., van Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H. H., Reitsma, J. B., and on behalf of the GetReal methods review group (2015). Get real in individual participant data (ipd) meta-analysis: a review of the methodology. *Research Synthesis Methods*, 6(4):293–309.

Jolani, S., Debray, T. P. A., Koffijberg, H., van Buuren, S., and Moons, K. G. M. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using mice. *Statistics in Medicine*, 34(11):1841–1863.

Quartagno, M. and Carpenter, J. R. (2016). Multiple imputation for ipd meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine*, 35(17):2938–2954.

Resche-Rigon, M. and White, I. R. (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, 27(6):1634–1649. PMID: 27647809.

Steyerberg, E. and Harrell, F. (2015). Prediction models need appropriate internal, internal-external, and external validation. *Journal of Clinical Epidemiology*, 69:245–247.