Incorporating published univariable associations in diagnostic and prognostic modeling

Thomas Debray

Julius Center for Health Sciences and Primary Care
University Medical Center Utrecht
The Netherlands

December 12, 2011

University Medical Center
Utrecht

Julius Center
for Health Sciences and Primary Care

Clinical Prediction Modeling

Aim

- provide a probability of outcome presence (diagnosis) or occurrence (prognosis) in an individual

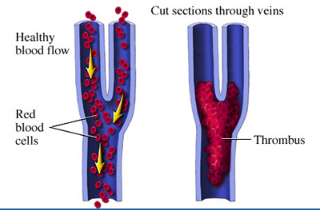
Typical Approach

1. Collection of Individual Patient Data (IPD)
2. Data Analysis (descriptives, missing values, ...)
3. Investigation of potential predictors
4. (Logistic) Regression Modeling
5. Evaluation of generalizability: validation studies

Practical Example

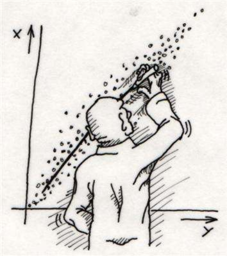
Diagnosis of Deep Vein Thrombosis

- Derivation dataset (IPD) of 1,295 patients
- Predictors: gender, oral contraceptive use, presence of malignancy, recent surgery, absence of leg trauma, vein distension, calf difference, D-dimer test
- Logistic Regression Modeling
- Validation dataset of 1,756 patients
  - Discrimination: 0.86 (AUC)
  - Calibration: 1.12 (Calibration slope)



Improving Generalization

- Increase Sample Size
  - Individual Participant Data
  - Individual Study Centers
- Amplify Sample Spectrum
  - Domain
  - Heterogeneity
- Apply Robust Estimation
  - Penalization & Shrinkage
  - Model Updating
  - Including External Knowledge



The Adaptation Method

- Introduced by Steyerberg/Greenland
- Re-estimates a multivariable coefficient
- Incorporates univariable coefficients from literature (e.g. log odds ratios for binary outcomes)

$$\beta_{\mathrm{m|L}} = \beta_{\mathrm{u|L}} + \left(\beta_{\mathrm{m|I}} - \beta_{\mathrm{u|I}}\right)$$

$$\mathrm{var}\left(\beta_{\mathrm{m|L}}\right) = \mathrm{var}\left(\beta_{\mathrm{u|L}}\right) + \mathrm{var}\left(\beta_{\mathrm{m|I}}\right) - \mathrm{var}\left(\beta_{\mathrm{u|I}}\right)$$

The Improved Adaptation Method

- Unbiased variance component

$$\mathrm{var}\left(\beta_{\mathrm{m|L}}\right) = \mathrm{var}\left(\beta_{\mathrm{u|L}}\right) + \mathrm{var}\left(\beta_{\mathrm{m|I}}\right) + \mathrm{var}\left(\beta_{\mathrm{u|I}}\right) - 2\mathrm{cov}\left(\beta_{\mathrm{m|I}}, \beta_{\mathrm{u|I}}\right)$$

- Distributional

$$\beta_{\mathrm{u|L}} \sim \mathcal{N}\left(\mu_{\mathrm{u|L}}, \sigma^2_{\mathrm{u|L}}\right), \beta_{\mathrm{m|I}} \sim \mathcal{N}\left(\mu_{\mathrm{m|I}}, \sigma^2_{\mathrm{m|I}}\right), \beta_{\mathrm{u|I}} \sim \mathcal{N}\left(\mu_{\mathrm{u|I}}, \sigma^2_{\mathrm{u|I}}\right)$$

- Robust Estimation

$$\mu_{\mathrm{m|I}} \sim \mathrm{Cauchy}\left(0, 2.5\right), \mu_{\mathrm{u|I}} \sim \mathrm{Cauchy}\left(0, 2.5\right)$$

Performance study

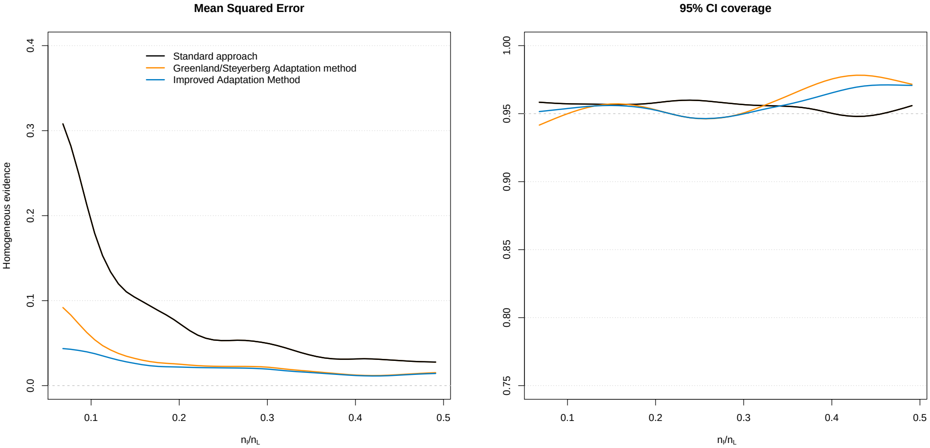
Simulation study

- Reference model with 2 predictors for generating data with $x_1, x_2 \sim \mathcal{N}(0,1)$ and $r(x_1, x_2) = 0$
- Individual Patient Data ($n_{\mathrm{IPD}} = 100 \rightarrow 1000$)
- 4 heterogeneous literature studies ($n_j = 500$)

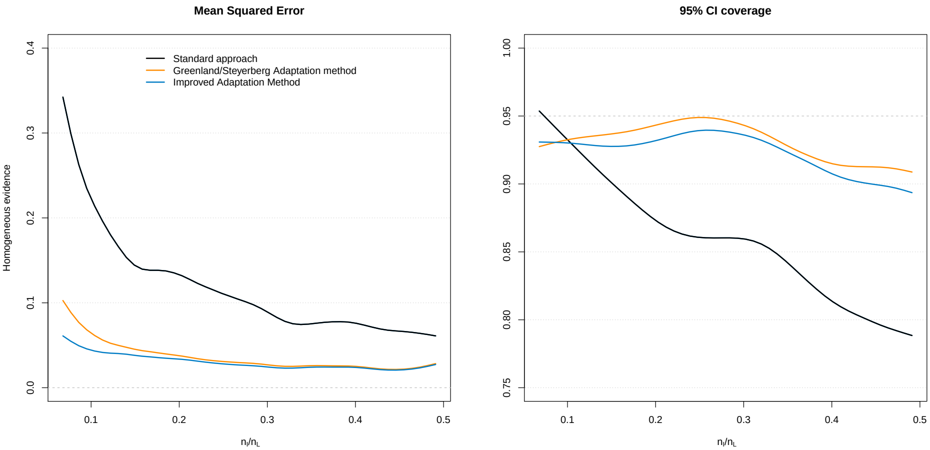
Case study: Diagnosis of Deep Vein Thrombosis

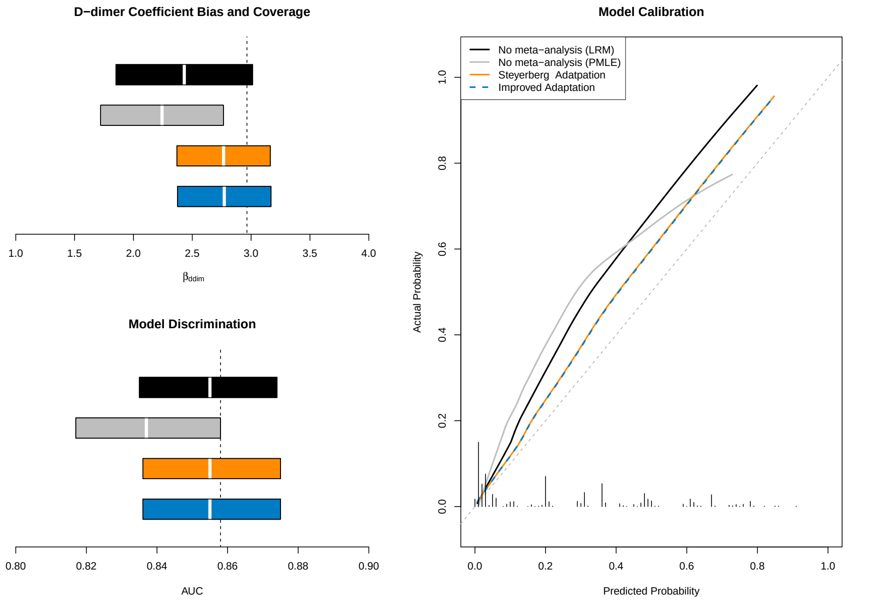
- IPD: Multivariable dataset ($n = 1,295$)
- LIT: 7 unadjusted odds ratios (biomarker D-dimer)
- Update D-dimer coefficient in multivariable prediction model
- External validation of updated prediction model ($n = 1,756$)

Simulation Study: homogeneous literature evidence



Simulation Study: heterogeneous literature evidence




D−dimer Coefficient Bias and Coverage
1.0
1.5
2.0
2.5
3.0
3.5
4.0
β_{ddim}
Model Discrimination
0.80
0.82
0.84
0.86
0.88
0.90
AUC
Model Calibration
No meta−analysis (LRM)
No meta−analysis (PMLE)
Steyerberg Adatpation
Improved Adaptation
Actual Probability
0.0
0.2
0.4
0.6
0.8
1.0
Predicted Probability

Discussion

- Strengths
  - Aggregation usually improves estimation
  - Abundance of external knowledge
  - Straightforward implementation of approaches
  - Explicit aggregated models (no black boxes)
- Weaknesses
  - Heterogeneity of external knowledge
  - Performance gain not always very large
  - Additional efforts required during derivation phase
- Ongoing research
  - incorporation of previously published prediction models with similar and different predictors