Chapter 14
Estimation with External Information

Background In this chapter we discuss methods that estimate regression coefficients based on the combination of findings from the sample under study with external information. We start with a simple "adaptation" method for univariate regression coefficients, which may be obtained from meta-analysis. This method was applied in a case study of operative mortality of abdominal aneurysm surgery. Next, we discuss some alternative approaches to estimate regression coefficients, including Bayesian estimation with explicit prior information.

14.1 Combining Literature and Individual Patient Data

We consider the common situation that several studies have already have been published for a particular clinical prediction problem, in which the relation between patient characteristics and the outcome of interest is described. If the published papers describe comparable patient series, we may try to combine the available evidence quantitatively in a meta-analysis. The information in these papers is usually only sufficient to calculate a univariate regression coefficient for each of the patient characteristics.

Multivariable coefficients can directly be estimated if individual patient data are available from the published series, or if we know the correlation structure between predictors. This information is usually not available. Individual patient data may be especially hard to retrieve for papers published several years ago, and anyway requires a substantial research effort. Thus, typically the researcher may have access to individual patient data from one study ("own data set") and univariate information from the literature ("publicly available").

An "adaptation method" has been proposed to take advantage of the univariate literature data in the estimation of the multivariable regression coefficients in a prediction model.[411] The aim is better prediction of the outcome in individual patients. This adaptation method is closely related to an earlier proposal by Greenland for meta-analysis.[151] For example, when studying the relation between coffee consumption and acute myocardial infarction, one study may have corrected the regression coefficient for a confounder (for example alcohol consumption), while other studies have not. Greenland proposed to use the change from unadjusted to adjusted regression coefficient to adapt the unadjusted coefficients in the latter studies.

E.W. Steyerberg, *Clinical Prediction Models*,
DOI:10.1007/978-0-387-77244-8_14, © Springer Science+Business Media, LLC 2009

14.1.1 Adaptation Method 1

In our case of regression analysis on literature and individual patient data, the formula reads like

$$\beta_{m|I+L}=\beta_{u|L}+(\beta_{m|I}-\beta_{u|I}),$$

where $\beta_{m|I+L}$ refers to the multivariable coefficient based on the combination of individual patient data and literature data (the "adapted coefficient"), $\beta_{u|L}$ is the univariate coefficient from a meta-analysis of the literature, and $\beta_{m|I} - \beta_{u|I}$ is the difference between multivariable and univariate coefficient in the individual patient data (the "adaptation factor"). Hence, we simply use the change from univariate to multivariable coefficient in our own data to adapt the meta-analysis coefficient.

For the variance of the adapted coefficient ($\text{var}(\beta_{m|I+L})$), we may add the difference between variances of the multivariable and univariate coefficient to the variance of the univariate coefficient from the literature, ignoring all covariances:

$$\text{var}(\beta_{m|I+L}) = \text{var}(\beta_{u|L}) + \text{var}(\beta_{m|I}) - \text{var}(\beta_{u|I}).$$

14.1.2 Adaptation Method 2

A more general way to formulate the adaptation formula is as

$$\beta_{m|I+L}=\beta_{m|I}+ c\,(\beta_{u|L}-\beta_{u|I}),$$

where c is a factor between 0 and 1. If $c = 1$, the same formula as proposed by Greenland arises. If c equals 0, the literature data is effectively discarded. The estimate of $\beta_{m|I+L}$ is unbiased for any choice of c, if the expectation of $\beta_{u|L} - \beta_{u|I} = 0$, that is, the individual patient data form a random part from the studies included in the meta-analysis. It was found that we can derive a formula for c so as to minimize the variance of $\beta_{m|I+L}$:

$$C_{\text{opt}} = \rho(\beta_{m|I}, \beta_{u|I}) \frac{\text{SE}(\beta_{m|I}) \times \text{SE}(\beta_{u|I})}{\text{var}(\beta_{u|L}) + \text{var}(\beta_{u|I})},$$

where $\rho(\beta_{m|I} - \beta_{u|I})$ refers to the correlation between multivariable and univariate coefficients in the individual patient data.

This variant of the adaptation method indicates that adaptation will be especially advantageous if the literature data set is larger (resulting in a smaller $\text{var}(\beta_{u|L})$), or when the correlation $\rho(\beta_{m|I} - \beta_{u|I})$ is larger. The latter correlation is expected to be large if the collinearity between covariables is small. The adaptation factor will then be close to 1, and method 1 may yield good results.

14.1.3 Estimation

Meta-analysis techniques may be used to estimate the univariate coefficients from the literature data. The literature data may include the individual patient data for maximal efficiency. The meta-analysis may assume fixed effects (for example, Mantel-Haenszel method, or conditional logistic regression), or random effects (for example, DerSimonian Laird method, or likelihood-based methods[97]). The calculations for method 1 use estimates that are readily available. For example, logistic regression analysis with standard maximum likelihood (ML) provides estimates of the univariate and multivariable coefficients in the individual patient data.

For the second method, the estimation of the optimal adaptation factor requires estimates of the variances of the regression coefficients, and an estimate of the correlation between univariate and multivariable coefficients. The latter correlation cannot easily be estimated with logistic regression methods. We therefore used bootstrap re-sampling to calculate the coefficients $\beta_{m|I}$ and $\beta_{u|I}$ repeatedly, and their correlation ρ.

14.1.4 Simulation Results

The adaptation method was tested in the GUSTO-I data.[411] First, we assessed the correlation between multivariable and univariate coefficients across 121 small subsamples. We observed a strong correlation for the combination of age and sex in a 2 predictor model (Fig. 14.1). Results were somewhat less favorable for predictors with stronger collinearity. For example, weight and height had a Pearson correlation coefficient of 0.54, and the correlation between their univariate and multivariable coefficients was 0.80 and 0.83 in a bivariate model respectively. Overall, the strong $r(\beta_{m|I} - \beta_{u|I})$ supports the use of the adaptation method in medical data.

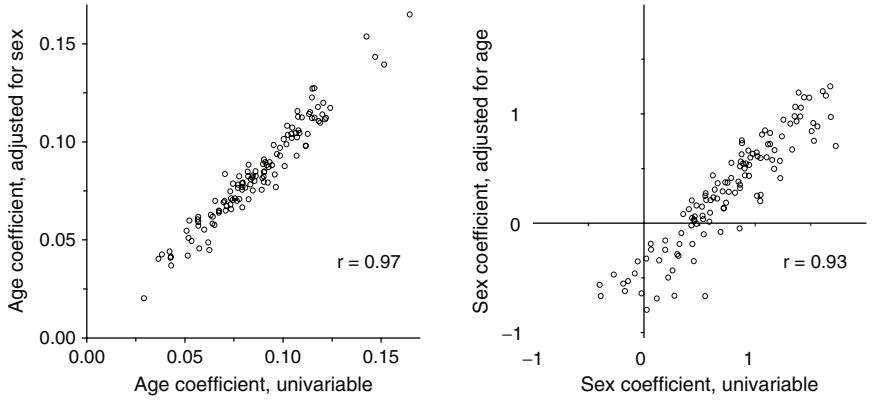

Fig. 14.1 Correlations between univariate and multivariable regression coefficients in a 2 predictor model consisting of age and sex estimated in 121 small subsamples of the GUSTO-I data set[411]

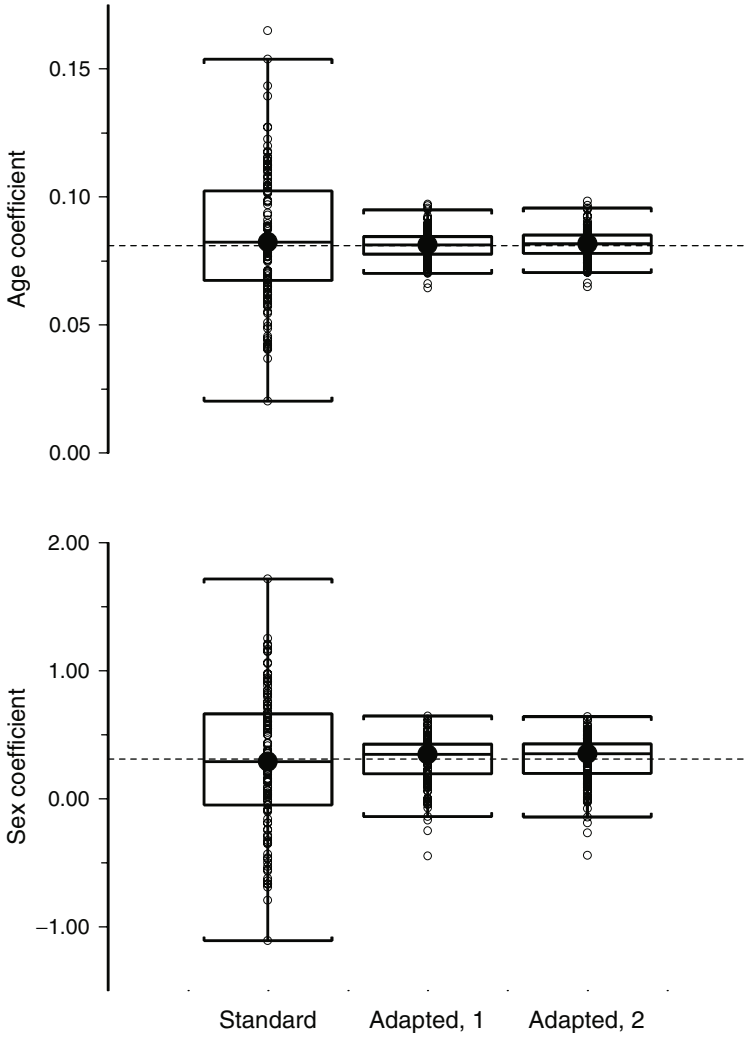

Fig. 14.2 Regression coefficients in the 2 predictor model consisting of age and sex. Box plots show the standard ML, and adapted estimates (methods 1 and 2) for 121 small subsamples; the line --- indicates the coefficient observed in the total GUSTO-I data set (n = 40,830)[411]

Next, we estimated the values of c_{opt}. Values were quite close to 1 (0.98 ± 0.015 and 0.99 ± 0.020 for age and sex (mean ± SD) in the 121 small subsamples). Hence, Greenland's method (c = 1) and our method (c estimated with bootstrapping) resulted in very similar estimates of the adapted coefficients (Fig. 14.2). Both methods lead to much better estimates of the multivariable regression coefficients in the small subsamples. Specifically, a substantial reduction is noted in the variability

compared with the standard multivariable regression coefficients, i.e. var($\beta_{m|I+L}$) << var($\beta_{m|I}$). These very favorable results were obtained by using univariate results from approximately half of the GUSTO-I data (n = 20,000). We also examined the influence of the size of the literature data. We applied the adaptation methods in the small subsamples, where univariate literature estimates were obtained from a neighboring, small subsample. This resulted effectively in a doubling of the sample size. This pattern was also reflected in the values of the adaptation factor from method 2; close to 1 with n = 20,000 as literature data, around 0.50 with a neighbour subsample as literature data.

14.1.5 Performance of Adapted Model

Finally, we compared the predictive performance of the adaptation method to the performance obtained with uniform shrinkage, penalized ML, or the Lasso in 23 large subsamples from GUSTO-I (Table 14.1). The discriminative ability improved slightly, but some problems were noted in calibration. Miscalibration was less than for the standard ML estimates, but some form of shrinkage should actually have been built into the adaptation method.

**14.1.6 Improving Calibration*

To improve the calibration of the predictions resulting from applying the adaptation method, we considered two approaches. First, we shrunk the multivariable regression coefficients as estimated in the individual patient data. This approach was discarded

Table 14.1 Discrimination (c statistic) and calibration (calibration slope) of the 8 and 17 predictor models based on large subsamples (average n=892, respectively), and based on the total training part (n=20,512), as evaluated in the independent test part of GUSTO-I (n=20,318)

	c statistic		Calibration slope	
	8 predictors	17 predictors	8 predictors	17 predictors
Total training (n=20,512, 1,423 deaths)				
Standard ML	0.789	0.802	0.944	0.959
23 large subsamples (n=892, 52 deaths on average)				
Standard ML	0.78	0.78	0.86	0.76
Uniform shrinkage	0.78	0.78	0.97	0.95
Penalized ML	0.78	0.79	0.96	0.98
Lasso	0.78	0.78	1.01	0.93
Adapted 1	0.79	0.79	0.92	0.86
Adapted 2	0.79	0.79	0.92	0.86

Means are shown for two variants of the adaptation method and several other modern estimation methods (see Chap. 13)

because it led to better calibration (slope closer to 1), but a decrease in discriminative ability. The second approach was motivated by the observation that the miscalibration of the adapted estimates was approximately halfway that of shrunk estimates and the standard ML estimates. The proposed formula is

$$\beta_{m|I} = (1 + \text{shrinkage factor}) / 2[\beta_{m|I} + c(\beta_{u|L} - \beta_{u|I})]$$

where the shrinkage factor is the uniform shrinkage factor, either estimated with a heuristic formula, or by bootstrapping (see Chap. 13).

Evaluations of this correction with method 1 (c set to 1) or 2 (c estimated by bootstrapping) showed an improvement in calibration. Discriminative ability was identical to that without shrinkage, since the shrinkage did not affect the ordering of predictions.

14.2 Example: Mortality of Aneurysm Surgery

In our examples with GUSTO-I, no relevant differences were noted between adaptation methods 1 and 2. We applied adaptation methods 1 and 2 in the prediction of peri-operative mortality (in-hospital or within 30 days) after elective abdominal aortic aneurysm (AAA) surgery.[421] Individual patient data were available on a relatively small sample (246 patients, 18 deaths). Patients were operated on at the University Hospital Leiden between 1977 and 1988. Univariate literature data were available from 15 published series with 15,821 patients (1,153 deaths) in total. Predictors considered included age and sex, cardiac comorbidity (history of myocardial infarction (MI), congestive heart failure (CHF), and ischemia on the ECG), pulmonary comorbidity (COPD, emphysema or dyspnea), and renal comorbidity (elevated pre-operative creatinin level). These predictors were chosen since they were reported in at least two studies in the literature, and were also available in the Leiden data set.

14.2.1 Meta-Analysis

Univariate logistic regression coefficients were estimated both with fixed and random effects methods from the literature data. As expected, the estimates of the coefficients were very similar, but the SEs were somewhat larger with the random effect method (Table 14.2).

A number of practical issues merit discussion with respect to the meta-analysis of the literature data. First, definitions of predictors varied, especially for pulmonary and renal comorbidity. Despite these differences, it was considered reasonable to assume one single effect for each predictor across the studies (non-significant tests for heterogeneity of odds ratios, non-significant interaction terms between study and effect estimates in logistic regression).

Table 14.2 Meta-analysis results for operative mortality of elective aortic aneurysm surgery (coefficient (SE))

Predictor	Fixed effect	Random effect
Age (per decade)	0.79 (0.06)	0.79 (0.11)
Female sex	0.36 (0.08)	0.36 (0.18)
History of MI	1.03 (0.27)	1.03 (0.32)
Congestive heart failure	1.59 (0.33)	1.59 (0.41)
ECG: Ischaemia	1.52 (0.31)	1.51 (0.38)
Impaired renal function	1.32 (0.25)	1.30 (0.26)
Impaired pulmonary function	0.89 (0.23)	0.85 (0.24)

Table 14.3 Individual patient data results (n=246) for operative mortality of elective aortic aneurysm surgery (coefficient (SE))

Predictor	Univariate	Standard ML	Shrunk	Penalized	$r(\beta_{m\|I}, \beta_{u\|I})$
Age (per decade)	0.98 (0.38)	0.58 (0.39)	0.48	0.34	0.91
Female sex	0.28 (0.79)	0.30 (0.86)	0.25	0.17	0.81
History of MI	1.50 (0.50)	0.74 (0.57)	0.61	0.57	0.88
Congestive heart failure	1.78 (0.55)	1.04 (0.59)	0.86	0.67	0.92
ECG: Ischaemia	1.72 (0.55)	0.99 (0.62)	0.83	0.63	0.87
Impaired renal function	1.24 (0.70)	1.12 (0.77)	0.93	0.74	0.85
Impaired pulmonary function	0.84 (0.53)	0.61 (0.59)	0.51	0.39	0.90

Second, the number of studies that described a predictor varied. The effect of age was reported in 15 studies, sex and renal function in 6, pulmonary function in 5, MI in 3, and CHF and ECG findings in only 2 studies. This somewhat limits the value of the adaptation method in this case study.

Third, the analysis of age as a continuous variable was hampered by the fact that mortalities were described in relatively large age intervals, for example, younger or older than 70 years. For logistic regression analysis, we estimated the mean ages in these age intervals using study-specific descriptions as far as available (mean and SE). We checked in a small simulation study that using the mean was better than using the median for age categories. The effect of age would have been estimated more accurately if smaller age intervals had been reported or more study characteristics had been published.

14.2.2 Individual Patient Data Analysis

In the individual patient data, multivariable logistic regression coefficients were usually smaller than the univariate coefficients, reflecting a predominantly positive correlation between predictors (Table 14.3). Correlations were strongest between the three cardiac comorbidity factors (r, 0.26, 0.32, and 0.45) and between these three factors and age (r>0.20). We note that the number of predictors (7) was large

relative to the number of events (18 deaths). Bootstrapping estimated a shrinkage factor of 0.83 (200 replications, convergence in only 119), and penalized ML was performed with 14 as the penalty factor. The correlation ρ between univariate and multivariate coefficients was estimated between 0.81 and 0.91.

14.2.3 Adaptation Results

The literature and individual patient data were combined with the adaptation method, using the random effect estimates from the literature data. For adaptation method 1, c_{opt} was always set to 1 (Table 14.4; for method 2, c_{opt} was estimated between 0.63 and 0.86, results not shown). Compared with shrunk or penalized coefficients, the adapted estimates for sex and renal and pulmonary function were somewhat higher and lower for a history of MI.

For application in clinical practice, scores were created by rounding each adapted coefficient after multiplication by 10 and shrinkage of 90% ((1+bootstrap shrinkage factor)/2≈0.90). The intercept was calculated with an offset variable in a logistic regression model. The offset was the linear combination of the scores (divided by ten) and the values of the covariables in the individual patient data. The intercept was estimated as −3.48.

The intercept was further adjusted for a presumably lower mortality in current surgical practice (5%) than that observed in the individual patient data (7.6%). This adjustment can be considered as a form of recalibration to contemporary circumstances. It was achieved by subtracting ln(odds(5%)/odds(7.6%)=–0.44 from the previous intercept estimate: −3.48 − 0.44 = −3.92. This results in the following formula to estimate the risk of peri-operative mortality in current elective abdominal aortic aneurysm surgery:

$$\mathrm{p(operativemortality)} = \frac{1}{[1+\exp(-(\sum \mathrm{score}/10) - 3.92)]}.$$

The area under the ROC curve was 0.83 in the individual patient data with standard, shrunk or penalized estimation. But the optimism-corrected estimates were

Table 14.4 Individual patient data results (n=246) for operative mortality of elective aortic aneurysm surgery (coefficient (SE))

Predictor	$\beta_{m\|I} - \beta_{u\|I}$	c method 1	Adapted 1	Score
Age (per decade)	−0.40	1	0.38 (0.14)	3
Female sex	+0.02	1	0.38 (0.40)	3
History of MI	−0.76	1	0.27 (0.41)	2
Congestive heart failure	−0.74	1	0.85 (0.47)	8
ECG: Ischaemia	−0.73	1	0.79 (0.48)	7
Impaired renal function	−0.12	1	1.18 (0.41)	11
Impaired pulmonary function	−0.23	1	0.62 (0.34)	6

Score: Rounded value of 9 × “Adapted 1”

0.80 for standard or shrunk, estimation, and 0.81 for penalized estimation (bootstrapping with 200 replications). For the final model with adapted coefficients, we expect a performance at least as good as these methods, but this needs to be confirmed in further validation studies.

14.3 Alternative Approaches

Several alternative approaches are possible to adjust univariate results for use in a multivariable model. We discuss two approaches below: Using an overall calibration factor for the univariate literature coefficients and Bayesian methods.

14.3.1 Overall Calibration

One variant of naïve Bayes was already suggested in Chap. 4, i.e. use of a uniform, overall calibration factor for all univariate coefficients. In the case study of aortic aneurysm mortality, the calibration factor is 0.69 for a linear predictor based on the univariate coefficients from the literature multiplied with the predictor values in the individual patient data. The recalibrated coefficients are reasonably close to those estimated with our adaptation method. The overall calibration led to higher values for cardiac comorbidity factors (scores 7, 11, and 10 for MI, CHF, and Ischaemia vs. 2, 8, and 7 with the adaptation method, respectively). This is explained by the relatively strong correlations among these factors, while the overall calibration reflects an average correlation between all the seven predictors.

14.3.2 Bayesian Methods: Using Data Priors to Regression Modelling

Greenland has argued that a Bayesian perspective needs to be incorporated into basic biostatistical and epidemiological training.[152] In particular in small data sets with many predictors, Bayesian approaches may offer advantages over conventional frequentist methods. Estimation of regression coefficients is difficult for data sets with few or no subjects at crucial combinations of predictor values.

Bayesian estimation consists of setting prior values for the regression coefficients, which are combined with the estimates in the data to produce posterior estimates of the coefficients. When the prior values are all zero, the coefficients are pulled towards zero. This is similar to shrinkage, as discussed in Chap. 13. Setting a prior to zero may be reasonable for a variable with very doubtful value as a predictor. A negative or positive effect is then equally likely, making zero the best prior guess. We allow for the possibility that the effect is non-zero, but may consider

large values unlikely. The degree of shrinkage is then determined by the width of the prior distribution. The narrower the prior distribution, the more the prior shrinks the coefficient towards zero. The other factor determining shrinkage is how strongly the predictor is related to the outcome in the data under study; in an informative data set (many events, not a rare predictor), there will be limited shrinkage. The final estimate is an average of the prior expectation and the conventional estimate.

A more interesting role for Bayesian approaches in regression is in using informative priors. For example, we may hypothesize a priori that a predictor has an odds ratio of 2, with values smaller than 0.5 and larger than 8 being highly unlikely. Setting a reasonable informative prior is the most difficult task for Bayesian analysis. Expert judgment or literature review can be used. When using informative priors, the source of these priors should be well documented, and sufficient variability allowed in the prior distribution. Presentation of prior information can be presented as "informationally equivalent," e.g. assuming knowledge of 100 patients with a certain outcome. This may be acceptable to some in the medical field, but will be met with scepticism by others, including traditional biostatisticians and applied clinical researchers.

*14.3.3 Example: Predicting Neonatal Death

Greenland describes a case study of predicting neonatal-death risk in a cohort of 2,992 births with 17 deaths.[152] He estimates logistic regression models with 14 predictors, assuming small to large effects for most predictors. He finds that the predictive ability of the Bayesian model is better than a model based on standard ML. He also illustrates how Bayesian estimation can be achieved relatively easily with data augmentation: Records are added to a data set, reflecting predictive effects of predictors.[153] In the case of a multivariable model, the prior distributions refer to the multivariable effects of predictors, which may be more complicated to elicit from experts or from literature than univariate effects.

*14.3.4 Example: Mortality of Aneurysm Surgery

In the prediction of peri-operative mortality of aortic aneurysms, we might try to use informative priors based on the literature. The meta-analysis however provides univariate effects, and we need to translate these to priors for multivariable effects. The difference between univariate and multivariable coefficients is directly related to the correlation between predictors. If we have some guesses for these correlations, this may give some hints on how the multivariable coefficients compare with the univariate coefficients. For example, with substantial correlations, we might halve all univariate coefficients; with no correlation, we keep the multivariate effect at the univariate estimate. Being on the conservative side with informative priors may be sensible to make Bayesian analysis more acceptable.

14.4 Concluding Remarks

The proposed adaptation methods emphasize the central role of subject knowledge in developing prediction models in small data sets. Literature data may guide the selection of predictors (Chap. 11), as well as improve the estimates of the regression coefficients (this chapter). Especially when the data set is relatively small, this strategy will result in more reliable regression models than using a strategy that considers a data set with individual patient data as the sole source of information.

A potential problem of meta-analyses is that publication bias may have led to overestimation of the regression coefficients. Also, performing a meta-analysis may not be realistic if definitions of risk factors vary substantially in the literature. Finally, the central assumption in the adaptation method is that the data set under study and the literature data are random subsamples from a common population, which implies that the correlations between predictors are similar in the individual patient data and in the literature data.

Bayesian methods provide another perspective on estimation of regression coefficients. If no effect is expected for a predictor, shrinkage of coefficients towards zero is achieved, quite similar to using uniform shrinkage or penalized ML. If other effects are assumed, coefficients will be pulled towards this prior value. As with any Bayesian method, the main criticism will be on the choice of prior distribution.

Many papers have been written about Bayesian approaches, but Bayesian methods have not yet made it to mainstream predictive modelling. A variant is empirical Bayes estimation, which will be discussed in Chaps. 20 and 21. Empirical Bayes methods have an important role in for example estimating centre effects, and provider profiling. With this variant, the prior distribution of centre effects is determined empirically from the data.

In some Bayesian applications, uninformative priors are used by default; these variants only use Bayesian calculations to achieve results that are difficult to calculate with frequentist methods, such as ML. These methods are becoming quite popular in medicine, e.g. using WinBUGS (www.mrc-bsu.cam.ac.uk/bugs/) with the Gibbs sampler as the core Bayesian method.[134]

Questions

14.1 Key factors in adaptation method (Sect. 14.1 and 14.2)
We examine the key factors for the adaptation method, as illustrated in the aneurysm case study.
(a) What would happen to the adapted coefficients when larger univariate coefficients were found in the literature?
(b) What would happen to the adapted coefficients when the univariate coefficients were identical in the literature and in the individual patient data?
(c) What would happen to the adapted coefficients when there was virtually no correlation between predictors?

14.2 Variance of adapted coefficients (Sect. 14.1.1)
In the simple variant, the variance of the adaption method is estimated as:
$\text{var}(\beta_{m|I+L})=\text{var}(\beta_{u|L})+\text{var}(\beta_{m|I}) - \text{var}(\beta_{u|I})$
When we have a literature data base ("L") of the same size as the individual patient data base ("I"), the variance decreased by a factor of 2 (SE decreases by $1/\sqrt{2}$, Sect. 14.1.4). What may be expected for the variance and SE of an adapted coefficient when we have a literature data base of 3 times the size of the individual patient data?

14.3 Adaptation method in aneurysm case study (Sect. 14.2)
For the aneurysm case study, the age effect is based on a very large sample size in the meta-analysis. The regression coefficient is 0.79 per 10 years; SE in random effect model, 0.14.
(a) Verify that the adaptation factor $\beta_{m|I} - \beta_{u|I}$ is −0.40.
(b) Verify that the SE of the adapted coefficient becomes 0.14, while it was 0.39 in the original multivariable analysis (Table 14.3).