STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:1965–1982
Published online 1 April 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2078

Cost–effectiveness analysis for multinational clinical trials

Eleanor M. Pinto[1,2], Andrew R. Willan[1,2,*,†] and Bernie J. O'Brien[3,4,✠]

[1]*Program in Population Health Sciences, Research Institute, Hospital for Sick Children, 555 University Avenue, Toronto, ON, Canada M5G 1X8*
[2]*Department of Public Health Sciences, University of Toronto, Toronto, ON, Canada*
[3]*Program for Assessment of Technology in Health, St Joseph's Healthcare, Hamilton, ON, Canada*
[4]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada*

SUMMARY

Clinical trials of cost–effectiveness are often conducted in more than one country. The two most common ways of dealing with the multinational nature of the data are either to calculate a pooled estimate or to stratify results by country. Since the between-country heterogeneity in costs is potentially substantial, pooled estimates may be difficult to interpret for any one country. Policy decisions are often made at a national level, and so country-specific results are important. However, country-specific analyses will be based on fewer patients and will often fail to provide adequate precision for statistical analyses.

Shrinkage estimation is a compromise between these two methods and has been used successfully in other fields. These estimates are country-specific yet less variable than those derived through a subgroup approach. Univariate and multivariate shrinkage estimators for costs and effects are proposed, then compared with one another and to the traditional methods in a simulation study. The methods are illustrated using data from a multinational trial evaluating the cost–effectiveness of three thrombolytic drug regimens in patients with acute myocardial infarction. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: multinational clinical trials; cost–effectiveness; shrinkage estimators

1. INTRODUCTION

The increasingly common practice of collecting individual patient cost data in randomized clinical trials has motivated the development of new statistical methodology designed to answer questions regarding the cost–effectiveness of new therapies. Early development focused

*Correspondence to: Andrew R. Willan, Population Health Sciences, Hospital for Sick Children, 555 University Avenue, Toronto, ON, Canada M5G 1X8.
†E-mail: andy@andywillan.com
✠Deceased

Contract/grant sponsor: Aventis
Contract/grant sponsor: Canadian Institutes of Health Research
Contract/grant sponsor: Ontario Student Opportunity Trust Fund-Hospital for Sick Children Foundation Student Scholarship Programme
Contract/grant sponsor: Discovery Grant Programme of the Natural Science and Engineering Research Council of Canada; contract/grant number: 44868-03

Received March 2004
Accepted September 2004

on making inference related to the incremental cost–effectiveness ratio (ICER) [1–10]. More recently, due to the problems associated with ratio statistics, attention has shifted to incremental net benefit (INB) [11–23]. Cost–effectiveness trials are often multinational, affording greater statistical power from the resulting increase in sample size. Other advantages include the perception of greater generalizability and the opportunity for the sponsor, in the case of a drug trial, to use the results for registration in more than one country.

The strength of these advantages relies to some degree on homogeneity of treatment differences across countries, allowing results to be pooled. While a strict protocol may increase homogeneity in effectiveness, the same cannot always be said for costs, since much of the heterogeneity arises at the national rather than the patient level. In the past the analysis of effectiveness trials has often been done without regard for the multinational nature of the data. (At best, analysts have added a fixed effect for country.) This reflects an underlying belief that treatment effectiveness is blind to international borders, and that investigators expect the estimate of treatment effect, and its corresponding uncertainty, to apply to the patients in their practice. While physicians may make treatment decisions at the patient level based on evidence of effectiveness, there may be greater reluctance to base treatment policy decisions on evidence of cost–effectiveness from the pooled results of multinational trials. This reflects the belief that there is more heterogeneity between countries with respect to costs and the fact that treatment policy decisions based on evidence of cost–effectiveness are usually made at the jurisdiction (often country) level by health policy makers rather than physicians. Consequently, policy makers often require 'country-specific' results.

Cook *et al.* [24] demonstrate the use of statistical tests for country by treatment interactions for various outcomes, including effectiveness, cost, INB and ICER, and suggest pooling across countries in the absence of heterogeneity. Similarly, regression methods [25, 26] rely on testing for interaction, although some of the heterogeneity can be removed by including covariates.

Multilevel models in which the between-country variance is incorporated through a random intercept provide an alternative, see Reference [27]. These models yield an overall pooled estimate of the treatment difference. In addition, the intercept estimates provide country-specific estimates. In contrast to the either/or approach offered by interaction testing, these estimates are a compromise, being a weighted average of the overall pooled estimate and the estimates based on country-specific data. The intercept estimates were first proposed by James and Stein [28] in the context of Bayesian decision theory in response to the observation that the maximum likelihood estimates based on the country-specific data are inadmissible under squared-error loss. These so-called 'shrinkage' estimators were later shown to be equivalent to the parametric empirical Bayes estimators in the case where the observations and intercept have normal distributions, see Reference [29].

The term shrinkage estimate is used because the estimate will be closer (i.e. shrunken) to the overall pooled estimate. The amount of shrinkage depends on the relative size of the between- and within-country variances; the smaller the between-country variance, in relation to the within-country variance, the more the estimate is shrunken to the overall estimate. The appeal of the shrinkage estimator is that it borrows strength from across countries. Willan *et al.* [30] discuss the use of shrinkage estimators for estimating country-specific between-treatment differences in mean cost for multinational trials. In this paper, we propose the use of shrinkage estimates for both cost and effectiveness to provide a more efficient country-specific cost–effectiveness analysis than using the country-specific data alone. The estimation problem is essentially one of estimating the between- and within-country covariance matrices.

The within-country covariance matrix can be estimated through the second moments based on the patient-level data. Provided that the within-country treatment differences in cost and effectiveness come from the same distribution, it is not difficult to obtain unbiased estimators for the between-country covariance matrix. If, however, there are important country-level covariates affecting the between-treatment differences, a regression approach is appropriate, and between-country covariance estimation becomes more complex. The problem is similar to multivariate random effects meta-analysis. For the multilevel model with one level for patient and another for country, it is identical.

In this paper, we use simulation analyses to compare a number of shrinkage estimators for cost and effectiveness with respect to efficiency and robustness to departures from the assumed nature of the random processes generating the data. The methods are illustrated using data from ASSENT-3, which is a multinational trial evaluating the relative cost–effectiveness of three thrombolytic drug regimens in patients with acute myocardial infarction.

2. METHODS

2.1. The model

Consider a multinational clinical trial with M countries in which patients are randomized between treatment (T) and standard (S). Let Δ_{ei} and Δ_{ci} be the differences between treatment means (T − S) for country i in effectiveness and cost, respectively. To carry out a country-specific cost–effectiveness analysis, using either an ICER or INB approach, Δ_{ei} and Δ_{ci} need to be estimated, along with corresponding variances and covariance. Let $\theta_i = (\Delta_{ei}, \Delta_{ci})^{\mathrm{T}}$ and $\hat{\theta}_i$ be an estimator of θ_i based on the data from country i alone. Let $V(\hat{\theta}_i) = V_i$, where $V(\cdot)$ is the variance–covariance function. Furthermore, suppose that $\hat{\theta}_i|(\theta_i, V_i) \sim \mathrm{N}(\theta_i, V_i)$ and that $\theta_i|(\tau, U) \sim \mathrm{N}(\tau, U)$. $\mathrm{N}(\tau, U)$ is the prior distribution for θ_i, and given $\hat{\theta}_i, \tau, V_i$ and U, the posterior distribution for θ_i is

$$\mathrm{N}(\tilde{\theta}_i, (V_i^{-1} + U^{-1})^{-1})$$

where $\tilde{\theta}_i = \tilde{\theta}_i(V_i, U, \tau) = (V_i^{-1} + U^{-1})^{-1}(V_i^{-1}\hat{\theta}_i + U^{-1}\tau)$. In practice, V_i, U and τ are unknown and are replaced by frequentist estimators $\hat{V}_i, \hat{U}$ and $\hat{\tau}$. Then, using the identities $(V_i^{-1} + U^{-1})^{-1} = V_i - V_i(U + V_i)^{-1}V_i$ and $(V_i^{-1} + U^{-1})^{-1}U^{-1} = V_i(U + V_i)^{-1}$, $\tilde{\theta}_i$ can be estimated by

$$\hat{\tilde{\theta}}_i = \hat{\theta}_i + \frac{M - K - R - 1}{M - R}\hat{V}_i(\hat{U} + \hat{V}_i)^{-1}(\hat{\tau} - \hat{\theta}_i) \tag{1}$$

where $R\ (=1)$ is one plus the number of covariates included in the prior for $\theta_i, K(=2)$ is the dimension of θ_i and the factor $(M - K - R - 1)/(M - R)$ corrects the bias from using $(\hat{U} + \hat{V}_i)^{-1}$ as an estimator for $(U + V_i)^{-1}$ (see Reference [31]).

Given V_i and U, the minimum variance, unbiased estimator of τ is

$$\hat{\tau} = W^{*-1}\sum_i W_i^*\hat{\theta}_i$$

where $W_i^* = (V_i + U)^{-1}$ and $W^* = \sum_i W_i^*$. The estimator $\hat{\tau}$, with variance W^{*-1}, is equivalent to the random effects meta-analysis estimator, pooling across countries. In practice, τ is estimated by $\hat{W}^{*-1}\sum \hat{W}_i^*\hat{\theta}_i$ and $V(\hat{\tau})$ by $\hat{W}^{*-1}$, where $\hat{W}_i^* = (\hat{V}_i + \hat{U})^{-1}$ and $\hat{W}^* = \sum_i \hat{W}_i^*$.

Since

$$\begin{aligned}\mathrm{Var}(\theta_i|\hat{\theta}) &= E(V(\theta_i|\hat{\theta},\tau,V_i,U)|\hat{\theta}) + V(E(\theta_i|\hat{\theta},\tau,V_i,U)|\hat{\theta})\\ &\approx V(\theta_i|\hat{\theta},\tau,V_i,U) + V(\tilde{\theta}_i|\hat{\theta})\\ &= V(\theta_i|\hat{\theta},\tau,V_i,U) + E(V(\tilde{\theta}_i(V_i,U,\tau)|\hat{\theta},U)|\hat{\theta}) + V(E(\tilde{\theta}_i(V_i,U,\tau)|\hat{\theta},U)|\hat{\theta})\\ &\approx V(\theta_i|\hat{\theta},\tau,V_i,U) + V(\tilde{\theta}_i(V_i,U,\tau)|\hat{\theta},U) + V(\tilde{\theta}_i(V_i,U,\hat{\tau})|\hat{\theta})\end{aligned}$$

the posterior variance of θ_i can be approximated by

$$V(\theta_i|\hat{\theta}) \approx V(\theta_i|\hat{\theta},\tau,V_i,U) + V(\tilde{\theta}_i(V_i,U,\hat{\tau})|\hat{\theta},U) + V(\tilde{\theta}_i(\hat{V}_i,U,\hat{\tau})|\hat{\theta}) \tag{2}$$

where $\hat{\theta} = \{\hat{\theta}_i, i = 1,\ldots,M\}$ and we note that the estimators $\hat{V}_i, \hat{U}, \hat{\tau}$ can be thought of as the posterior estimates of V_i, U and τ when these are given vague hyperpriors. Then, following Morris [32], the posterior variance of θ_i given the data and all the prior parameters is given by

$$V(\theta_i|\hat{\theta},\tau,V_i,U) = V_i - V_i(U+V_i)^{-1}V_i \approx \hat{V}_i - \frac{M-K-R-1}{M-R}\hat{V}_i(\hat{U}+\hat{V}_i)^{-1}\hat{V}_i \tag{3}$$

the added variability due to estimation of τ is given by

$$\begin{aligned}V(\tilde{\theta}_i(V_i,U,\tau)|\hat{\theta},U) &= V_i(U+V_i)^{-1}V(\tau|\hat{\theta},U)(U+V_i)^{-1}V_i\\ &\approx P_i\hat{W}^{*-1}P_i^{\mathrm{T}}\end{aligned} \tag{4}$$

with $P_i = (M-K-R-1)\hat{V}_i(\hat{U}+\hat{V}_i)^{-1}/(M-R)$; the added variability due to estimation of U is given by

$$V(\tilde{\theta}_i(\hat{V}_i,U,\hat{\tau})|\hat{\theta}) = V(V_i(U+V_i)^{-1}(\hat{\theta}_i-\hat{\tau})|\hat{\theta}) \approx \mathrm{diag}(v_i \times (\hat{\theta}_i-\hat{\tau}) \times (\hat{\theta}_i-\hat{\tau})) \tag{5}$$

where $\times$ denotes component-wise multiplication of vectors, v_i is a vector of length K with kth component

$$v_{i\{k\}} \approx V[(V_{i\{kk\}}+\hat{U}_{\{kk\}})^{-1}V_{i\{kk\}}] \approx \frac{2(M-K-R-1)^2(\hat{V}_{i\{kk\}})^2(\bar{V}_{\{kk\}}+\hat{U}_{\{kk\}})}{(M-R)^2(M-R-2)(\hat{V}_{i\{kk\}}+\hat{U}_{\{kk\}})^3}$$

$A_{\{ij\}}$ is the ith–jth element of the matrix A and the definition of $\bar{V}_{\{kk\}}$ depends on the estimator used for estimating U, as described in Section 2.2. The V_i can be estimated from the first and second moments of the patient-level, country-specific data or, in the case of censored data, using methods proposed by Willan and Lin [21] and Willan *et al.* [22, 23]. Possible estimators for U are introduced in the next section.

2.2. Estimators for U

Three possible estimators of U are proposed.

2.2.1. Unweighted.

$$\hat{U}_1 = \frac{\sum(\hat{\theta}_i - \hat{\tau}_1)(\hat{\theta}_i - \hat{\tau}_1)^{\mathrm{T}} - (1 - M^{-1})\sum_i \hat{V}_i}{M - 1}$$

where $\hat{\tau}_1 = \sum_i \hat{\theta}_i/M$. $\hat{U}_1$ is unbiased. For $\hat{U}_1$, $\bar{V}_{\{kk\}} = \sum_i V_{i\{kk\}}/M$.

2.2.2. Weighted.

$$\hat{U}_{2\{kl\}} = \frac{[\sum_i \hat{W}_i^{1/2}(\hat{\theta}_i - \hat{\tau}_2)(\hat{\theta}_i - \hat{\tau}_2)^{\mathrm{T}}\hat{W}_i^{1/2}]_{kl} - c_{kl}}{m_{kl}}$$

where $\hat{W}_i$ is a diagonal matrix, with jth diagonal element

$$\hat{W}_{i\{jj\}} = \hat{w}_{ij} = \frac{(\hat{V}_{i\{jj\}})^{-1}}{\sum_k (\hat{V}_{k\{jj\}})^{-1}}$$

$$\hat{\tau}_2 = \sum_i \hat{W}_i \hat{\theta}_i$$

$$c_{k\ell} = \sum_i \hat{w}_{ik}^{1/2}\hat{w}_{i\ell}^{1/2}\hat{V}_{i\{k\ell\}} + \sum_i \hat{w}_{ik}^{1/2}\hat{w}_{i\ell}^{1/2}\sum_j \hat{w}_{jk}\hat{w}_{j\ell}\hat{V}_{j\{k\ell\}} - \sum_i \hat{w}_{ik}^{3/2}\hat{w}_{i\ell}^{1/2}\hat{V}_{i\{k\ell\}} - \sum_i \hat{w}_{ik}^{1/2}\hat{w}_{i\ell}^{3/2}\hat{V}_{i\{k\ell\}}$$

and

$$m_{k\ell} = \sum_i \hat{w}_{ik}^{1/2}\hat{w}_{i\ell}^{1/2} + \sum_i \hat{w}_{ik}^{1/2}\hat{w}_{i\ell}^{1/2}\sum_j \hat{w}_{jk}\hat{w}_{j\ell} - \sum_i \hat{w}_{ik}^{3/2}\hat{w}_{i\ell}^{1/2} - \sum_i \hat{w}_{ik}^{1/2}\hat{w}_{i\ell}^{3/2}$$

$\hat{U}_2$ is unbiased. For $\hat{U}_2$, $\bar{V}_{\{kk\}} = M(\sum_i (\hat{V}_{i\{kk\}})^{-1})^{-1}$. The diagonal elements $\hat{U}_{2\{kk\}}$ are equal to the estimator of between-country variance based on the method of moments for a random effects model, see Reference [33].

2.2.3. Restricted maximum likelihood. $\hat{U}_3$ is the restricted maximum likelihood estimator of U, achieved through a combination of the EM, with an Aitken acceleration, and the Newton–Raphson algorithms. For $\hat{U}_3$, $\bar{V}_{\{kk\}} = (\sum_i (\hat{W}^*_{i\{kk\}})^2)^{-1}\sum_i (\hat{W}^*_{i\{kk\}})^2 \hat{V}_{i\{kk\}}$.

2.3. Negative variance estimates

Except for $\hat{U}_3$, which is maximized over the space of non-negative definite matrices, the diagonal elements of the estimate of U can be negative. This implies that the observed between-country variance in one or more of the components of $\hat{\theta}_i$ is less than would be expected, given the observed variance among patients within a country. In the scalar case (i.e. shrinking cost and effectiveness individually) it has been shown that $\hat{\sigma}_k^{2+} = \max(\hat{\sigma}_k^2, 0)$ strictly dominates $\hat{\sigma}_k^2$ with respect to squared error loss, see Reference [28]. Therefore, we propose that if both diagonal elements are negative, then U is estimated by $\left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, and $\hat{\tilde{\theta}}_i = (K+1)\hat{\theta}_i/(M-R) + (M-K-R-1)\hat{\tau}/(M-R)$, with $\hat{V}(\tilde{\theta}_i)$ as in (2)–(5). Since $\hat{W}_i^* = \hat{V}_i^{-1}$ in this case, $\hat{\tau}$ becomes the fixed effects meta-analysis estimator of τ.

If just one of the diagonal elements of the estimate of U is negative, we propose that it and the off-diagonal elements be replaced by zero. Equation (1) for $\hat{\tilde{\theta}}_i$ continues to apply, as do equations (2)–(5) for the posterior variance.

2.4. Scalar shrinkage

An alternative to determining a shrinkage estimator for the vector θ_i is to determine the shrinkage estimator for the components individually, in which case the between-country variance for the kth component is estimated by $\hat{\sigma}^2_{\{k\}} = \hat{U}_{\{kk\}}$, with $\hat{U}$ any of the estimators in Section 2.2. The formulae in Section 2.1 become

$$\hat{\tilde{\theta}}_{i\{k\}} = \hat{\theta}_{i\{k\}} + \frac{M-R-2}{M-R} \frac{\hat{\sigma}^2_{i\{k\}}}{\hat{\sigma}^2_{i\{k\}} + \hat{\sigma}^2_{\{k\}}} (\hat{\tau}_{\{k\}} - \hat{\theta}_i)$$

where $\hat{\tau}_k = \hat{w}^{*-1} \sum_i \hat{w}^*_{ik} \hat{\theta}_{i\{k\}}$, $\hat{w}_{i\{k\}} = (\hat{\sigma}^2_{i\{k\}} + \hat{\sigma}^2_{\{k\}})^{-1}$, $\hat{w}^*_k = \sum_i \hat{w}^*_{ik}$ and $\hat{\sigma}^2_{i\{k\}} = \hat{V}_{i\{kk\}}$. The posterior variance of the resulting shrinkage estimates $(\tilde{\theta}_{i\{1\}}, \tilde{\theta}_{i\{2\}})^{\mathrm{T}}$ can be estimated by equations (2)–(5), but now with $K = 1$,

$$P_i = \begin{pmatrix} \dfrac{M-R-2}{M-R} \times \dfrac{\hat{\sigma}^2_{i\{1\}}}{\hat{\sigma}^2_{i\{1\}} + \hat{\sigma}^2_{\{1\}}} & 0 \\ 0 & \dfrac{M-R-2}{M-R} \times \dfrac{\hat{\sigma}^2_{i\{2\}}}{\hat{\sigma}^2_{i\{2\}} + \hat{\sigma}^2_{\{2\}}} \end{pmatrix}$$

and

$$V(\hat{\tau}) = \begin{pmatrix} \hat{w}^{*-1}_1 & \hat{w}^{*-1}_1 \hat{w}^{*-1}_2 \sum\limits_i \hat{w}^*_1 \hat{w}^*_2 \hat{V}_{i\{12\}} \\ \hat{w}^{*-1}_1 \hat{w}^{*-1}_2 \sum\limits_i \hat{w}^*_1 \hat{w}^*_2 \hat{V}_{i\{21\}} & \hat{w}^{*-1}_2 \end{pmatrix}$$

2.5. Regression models

In some instances there may be country-level covariates which reduce the between-country heterogeneity. These can be incorporated into the shrinkage procedure by re-formulating the model as

$$\hat{\theta}_i | \theta_i, V_i \sim \mathrm{N}(\theta_i, V_i)$$

$$\theta_i | \beta, U \sim \mathrm{N}(\tau_i, U) \quad \text{with}$$

$$\tau_i = \beta X_i, \quad \beta = \begin{pmatrix} \beta_{\mathrm{e}1} & \cdots & \beta_{\mathrm{e}R} \\ \beta_{\mathrm{c}1} & \cdots & \beta_{\mathrm{c}R} \end{pmatrix}, \quad X_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{iR} \end{pmatrix}$$

so that β is a matrix of regression coefficients, X_i is a vector of covariates for country i and X_{i1} would usually be equal to 1 for all i. Equation (1) now becomes

$$\hat{\hat{\theta}}_i = \hat{\theta}_i + \frac{M-K-R-1}{M-R}\hat{V}_i(\hat{U}+\hat{V}_i)^{-1}(\hat{\tau}_i - \hat{\theta}_i)$$

and similarly (2)–(5) continue to apply, but with $\hat{\tau}$ replaced by $\hat{\tau}_i$. When the ture regression equation includes more than just an intercept, the unweighted and weighted estimators for U are no longer unbiased, so restricted maximum likelihood (REML) estimation should be used. In practice, estimation procedures iterate between estimating β by maximum likelihood and U by restricted maximum likelihood.

3. EXAMPLE: THE ASSENT-3 TRIAL

The ASsessment of the Safety and Efficacy of New Thrombolytic Regimens (ASSENT)-3 trial has been described elsewhere in detail [34, 35]. Patients with ST elevation acute myocardial infarction were randomized between three treatment arms:

- Heparin: full-dose tenecteplase plus unfractionated heparin;
- Enoxaparin: full-dose tenecteplase plus enoxaparin;
- Abciximab: half-dose tenecteplase plus unfractionated heparin plus abciximab.

The enoxaparin and abciximab arms were compared to the heparin arm with respect to safety and effectiveness. A total of 6095 patients were enrolled at 575 sites in 26 countries, with the largest number of patients (975) from the U.S. and the smallest number (23) from Finland. A total of 435 patients were enrolled from Canada. The measure of effectiveness was freedom from death, in-hospital re-infarction and refractory ischaemia for 30 days. Both investigational regimens (the enoxaparin and abciximab arms) showed a statistically significant improvement in effectiveness compared to the heparin arm. Effectiveness was observed in 1721 of the 2036 (84.5 per cent) patients randomized to the heparin arm, vs 1804 of the 2037 (88.6 per cent) patients randomized to the enoxaparin arm and 1794 of the 2017 (88.9 per cent) patients randomized to the abciximab arm.

As part of the trial, resource use data were collected on length of hospital stay, study drug and dosage, cardiac catheterization, percutaneous coronary intervention (PCI) and coronary artery bypass graft (CABG) surgery during the index hospitalization and length of stay and PCI and CABG surgery use during repeat hospitalizations within 30 days.

For the purposes of this report Canadian price weights (e.g. drug prices) for resources consumed were used to measure total 30-day cost for each patient in the trial. Estimates of hospital costs were obtained from a large teaching hospital in Southern Ontario using the Ontario Case Costing system [36] and professional fees from the Ontario Schedule of Benefits [37]. Prices, all given in 2003 Canadian dollars, were assigned to each health care resource component available from the clinical ASSENT-3 database to calculate the total 30-day cost for each patient.

The pooled results, showing the estimated difference in effectiveness and cost for each investigational regimen compared to the heparin arm, are shown in Table I for all three methods of estimating U. Estimates for U, τ and $V(\hat{\tau})$ for the various methods are shown in Table I for the enoxaparin vs heparin (E–H) and abciximab vs heparin (A–H) comparisons.

Table I. Pooled Estimates.

	Enoxaparin vs heparin			Abciximab vs heparin		
	$\hat{U}$	$\hat{\tau}$	$\hat{V}(\hat{\tau})$	$\hat{U}$	$\hat{\tau}$	$\hat{V}(\hat{\tau})$
Unweighted						
Scalar		$\begin{pmatrix} 0.03407 \\ -55.36 \end{pmatrix}$	$\begin{pmatrix} 0.0001038 & -0.5718 \\ -0.5718 & 19772 \end{pmatrix}$		$\begin{pmatrix} 0.04209 \\ 912.6 \end{pmatrix}$	$\begin{pmatrix} 0.0002030 & -1.037 \\ -1.037 & 42237 \end{pmatrix}$
	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$			$\begin{pmatrix} 0.001801 & -6.983 \\ -6.983 & 401681 \end{pmatrix}$		
Vector		$\begin{pmatrix} 0.03204 \\ -57.02 \end{pmatrix}$	$\begin{pmatrix} 0.00009917 & -0.6113 \\ -0.6113 & 18757 \end{pmatrix}$		$\begin{pmatrix} 0.03928 \\ 921.5 \end{pmatrix}$	$\begin{pmatrix} 0.0001994 & -1.058 \\ -1.058 & 41345 \end{pmatrix}$
Weighted						
Scalar		$\begin{pmatrix} 0.03407 \\ -55.36 \end{pmatrix}$	$\begin{pmatrix} 0.0001038 & -0.5718 \\ -0.5718 & 19772 \end{pmatrix}$		$\begin{pmatrix} 0.03994 \\ 935.7 \end{pmatrix}$	$\begin{pmatrix} 0.0001240 & -0.6273 \\ -0.6273 & 28578 \end{pmatrix}$
	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$			$\begin{pmatrix} 0.0003038 & -0.005418 \\ -0.005418 & 135194 \end{pmatrix}$		
Vector		$\begin{pmatrix} 0.03204 \\ -57.02 \end{pmatrix}$	$\begin{pmatrix} 0.00009917 & -0.6113 \\ -0.6113 & 18757 \end{pmatrix}$		$\begin{pmatrix} 0.03741 \\ 942.7 \end{pmatrix}$	$\begin{pmatrix} 0.0001198 & -0.6627 \\ -0.6627 & 27269 \end{pmatrix}$
REML						
Scalar		$\begin{pmatrix} 0.03457 \\ -54.98 \end{pmatrix}$	$\begin{pmatrix} 0.0001204 & -0.5145 \\ -0.5145 & 20124 \end{pmatrix}$		$\begin{pmatrix} 0.03981 \\ 943.9 \end{pmatrix}$	$\begin{pmatrix} 0.0001185 & -0.4027 \\ -0.4027 & 24834 \end{pmatrix}$
	$\begin{pmatrix} 0.0002417 & 1.013 \\ 1.013 & 5236 \end{pmatrix}$			$\begin{pmatrix} 0.0002183 & 3.519 \\ 3.519 & 70033 \end{pmatrix}$		
Vector		$\begin{pmatrix} 0.03381 \\ -52.97 \end{pmatrix}$	$\begin{pmatrix} 0.0001151 & -0.5501 \\ -0.5501 & 19255 \end{pmatrix}$		$\begin{pmatrix} 0.03730 \\ 948.7 \end{pmatrix}$	$\begin{pmatrix} 0.0001124 & -0.4410 \\ -0.4410 & 23375 \end{pmatrix}$

For E–H the elements of $\hat{U}$ for weighted and unweighted are 0 and, consequently, $\hat{\tau}$ is the fixed effect estimate and will be the same for weighted and unweighted. The elements of the REML estimate of U are not 0, since they are constrained not to be. However, the REML estimate of τ, which in this case is a random-effects estimate, is very similar and, because $\hat{\tau}$ is a random effects estimate, the variance elements of $\hat{V}(\hat{\tau})$ will be larger than those for weighted and unweighted. For all three methods, the variance elements of $\hat{V}(\hat{\tau})$ are smaller for vector, compared to scalar estimation.

For the A–H comparison the elements of $\hat{U}$ are non-zero for all methods. The diagonal elements are smallest for REML and largest for unweighted. The estimates of τ are all very similar, and no consistent or material difference is seen between methods. The diagonal elements of $\hat{V}(\hat{\tau})$ are smallest for REML and largest for unweighted, which is consistent with our observation regarding $\hat{U}$. For all three methods, the variance elements of $\hat{V}(\hat{\tau})$ are smaller for vector, compared to scalar estimation.

Parameter estimates for Canada are shown in Table II. The pooled random effect estimates are given in the top panel. These, for obvious reasons, will have the smallest standard errors. However, as seen in the simulation results in the following section, confidence intervals based on these estimates and standard errors have very poor coverage for individual country parameters. Estimates based on the Canadian data alone are shown in the second panel. These, for obvious reasons, will have the largest standard errors, although the corresponding confidence intervals, as illustrated in the next section, will have appropriate coverage. The shrinkage estimates are shown in the bottom three panels. For each method of estimating U, for both treatment comparisons, and for both effectiveness and costs, the vector estimates have slightly larger standard errors. The 'weighted' estimates have smaller standard errors than the 'unweighted', and except for the effectiveness for the E–H comparison, the REML estimates have smaller standard errors than 'weighted'.

The Canadian cost–effectiveness acceptability curves for the abciximab vs heparin contrast are shown in Figure 1(a). The curve for the observed data is least steep, and therefore would provide the widest credible intervals. The curves for the shrinkage estimates show a consistent pattern, with REML providing the narrowest intervals, followed by weighted and then unweighted. The curves for the enoxaparin vs heparin contrast are shown in Figure 1(b). The shrinkage estimates show no consistent pattern, while the curve for the observed data has negative slope because of the large amount of probability in the south-west quadrant, i.e. where enoxaparin is less costly and less effective.

4. SIMULATION COMPARISONS

4.1. Simulated data

To determine and compare the properties of the shrinkage estimators using different estimators of U, a simulation analysis was performed. The abciximab and heparin groups of the ASSENT trial were used to estimate U (by the weighted method), τ and V_i. Under the assumption that the between-patient variation was equal for each country and treatment arm, this was estimated by $\bar{V} = \frac{1}{2} \times \sum_i \hat{V}_i n_i / 25$, where n_i is the average number of patients per treatment arm for country i. Each simulated data set drew a set of true country-specific means θ_i from a multivariate normal with mean τ and variance U. Patient data for each country was

Table II. Treatment contrasts for Canada.

	Enoxaparin vs heparin		Abciximab vs heparin	
Estimate (Std. error) [p-value] {95 per cent CI}	Effectiveness	Cost	Effectiveness	Cost
Pooled random effects (REML, vector)	0.03381 (0.01073) [0.001626] {0.01278, 0.05483}	−52.97 (138.8) [0.7026] {−324.9, 219.0}	0.03730 (0.01060) [0.0004352] {0.01652, 0.05808}	948.7 (152.9) [<0.0001] {649.0, 1248}
Canadian patients only	0.008773 (0.04067) [0.8292] {−0.07093, 0.08848}	−731.2 (615.1) [0.2346] {−1937, 474.5}	0.06032 (0.0374) [0.1068] {−0.01299, 0.1336}	1008 (663.9) [0.1289] {−293.1, 2309}
Unweighted				
Scalar	0.03196 (0.01948) [0.1010] {−0.00623, 0.07015}	−111.7 (381.6) [0.7698] {−859.7, 636.4}	0.05301 (0.02967) [0.07396] {−0.005137, 0.1112}	962.3 (489.3) [0.04921] {3.330, 1921}
Vector	0.02913 (0.02019) [0.1490] {−0.01043, 0.06870}	−141.3 (392.8) [0.7191] {−911.1, 628.6}	0.05231 (0.03008) [0.08205] {−0.006651, 0.1113}	962.5 (497.6) [0.05304] {−12.65, 1938}
Weighted				
Scalar	0.03196 (0.01737) [0.06579] {−0.002088, 0.06601}	−111.7 (306.1) [0.7153] {−711.7, 488.4}	0.04497 (0.02126) [0.03441] {0.003299, 0.08664}	957.3 (382.0) [0.01220] {208.7, 1706}
Vector	0.02913 (0.01856) [0.1166] {−0.007253, 0.06552}	−141.3 (324.0) [0.6627] {−776.2, 493.7}	0.04408 (0.02231) [0.04822] {0.0003436, 0.08782}	981.2 (396.9) [0.01343] {203.3, 1759}
REML				
Scalar	0.02940 (0.02113) [0.1640] {−0.01201, 0.07081}	−119.8 (281.8) [0.6708] {−672.2, 432.6}	0.04406 (0.01978) [0.02590] {0.005294, 0.08282}	957.3 (328.4) [0.003557] {313.6, 1601}
Vector	0.02512 (0.02210) [0.2557] {−0.01820, 0.06845}	−162.7 (302.6) [0.5907] {−755.8, 430.4}	0.04350 (0.02039) [0.03288] {0.003539, 0.08345}	1012 (339.2) [0.002853] {347.0, 1676}

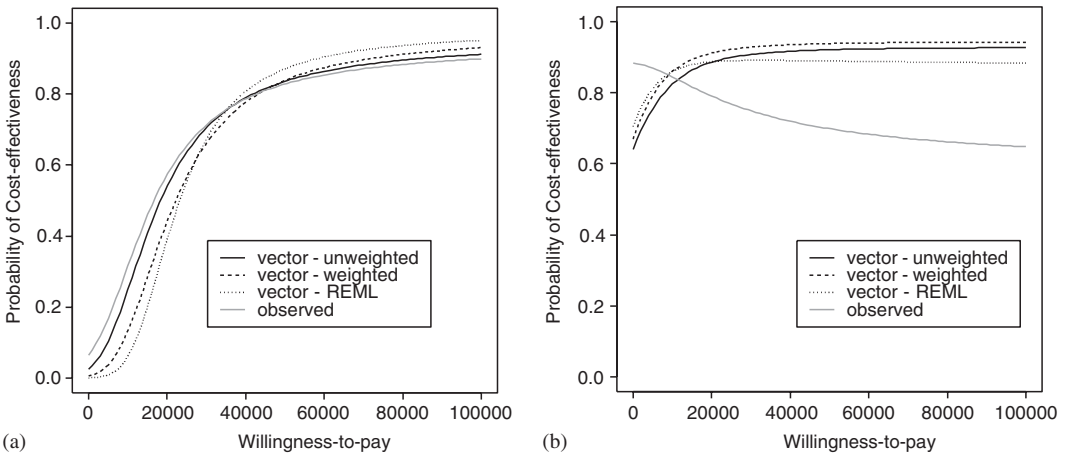


Figure 1. (a) Cost–effectiveness acceptability curves for abciximab vs heparin in Canada using the three vector shrinkage estimates and the observed country-specific estimate; and (b) cost–effectiveness acceptability curves for enoxaparim vs heparin in Canada using the three vector shrinkage estimates and the observed country-specific estimate.

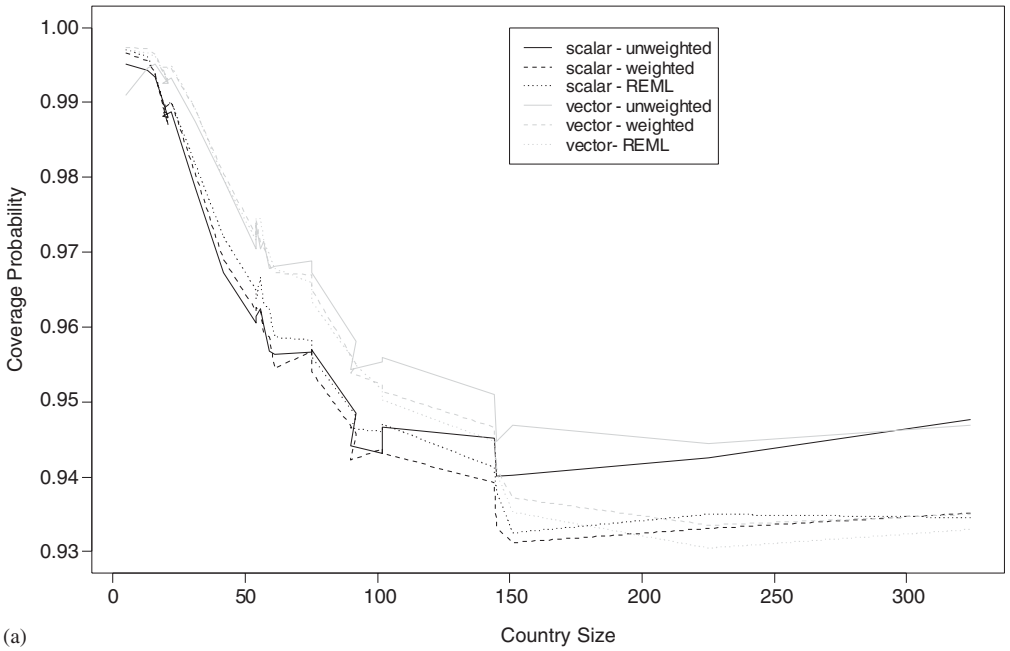
generated from a multivariate normal distribution with variance $\bar{V}$ and mean θ_i for those in the abciximab group and 0 for those in the heparin group. This was repeated 10 000 times. The average standard error of the shrinkage estimators was recorded, together with the proportion of times that the 95 per cent credible interval of the shrinkage estimate included the value of θ_i used to generate the data.

The sensitivity of the results to the assumed normal prior for the random effects was tested by repeating the simulation with random effects $\theta_i = (\theta_{i\{e\}}, \theta_{i\{c\}})^T$ generated from a pair uniform distributions, $U[0, m_e)$ and $U[0, m_c)$ with parameters m_e and m_c chosen so that the between-country variance in costs and effects was the same as for the normal simulation, although the between-country covariance was set at zero.

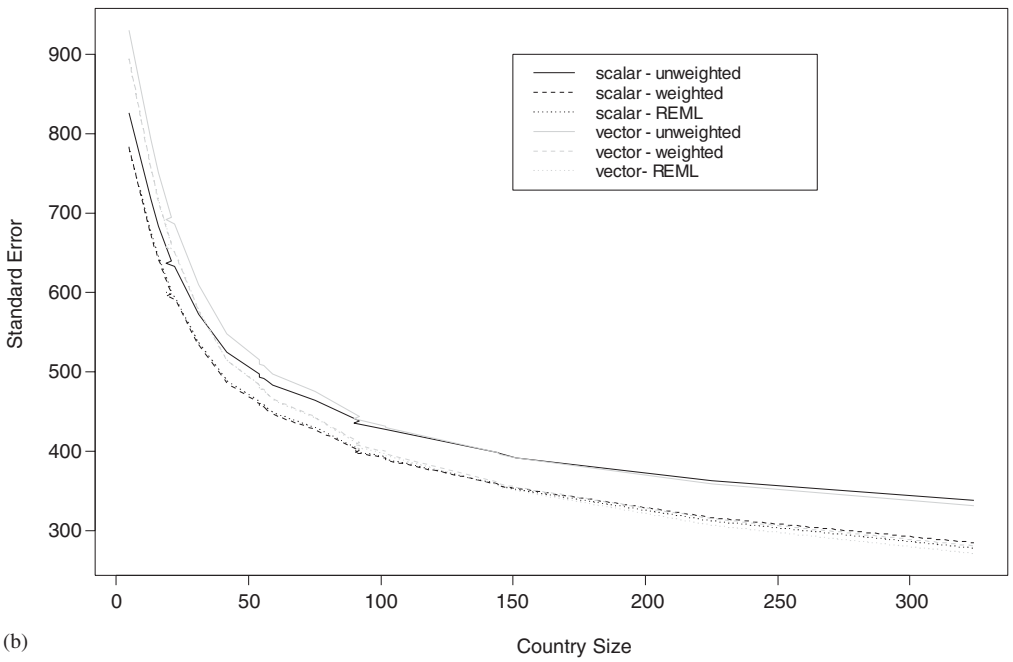
4.2. Simulation results

Figure 2(a) shows the estimated coverage probabilities for costs. Coverage probabilities for the shrinkage estimators depend on country sample size, exceeding 95 per cent in small to moderate sample sizes and falling slightly short at 93–94 per cent in the larger countries. The intervals based on the observed country-specific difference in means achieve almost 95 per cent coverage; the shortfall in smaller countries is due to ignoring the between-country variance when calculating standard errors. The pooled random effects estimator gives intervals with poor coverage for the θ_i, at 58–64 per cent depending on the choice of estimation procedure and country sample size.

The average standard errors for each country are shown in Figure 2(b). Amongst the shrinkage estimators, those derived using unweighted estimators for the between-country variance have the largest standard errors for moderate-to-large country sample sizes, with correspondingly higher coverage probabilities. Weighted and REML estimators perform very similarly.



(a)



(b)

Figure 2. (a) Coverage probabilities for cost by number of patients per treatment arm; and (b) cost average standard errors by number of patients per treatment arm.

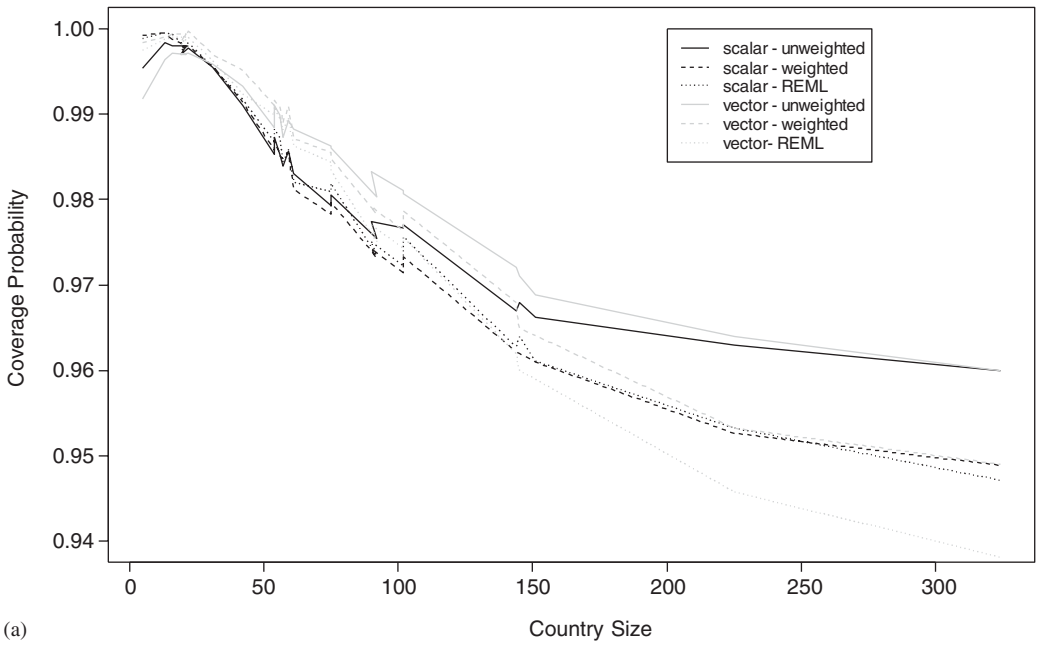
In small to moderate country sample sizes, vector shrinkage results in both larger standard errors and higher coverage, whilst the reverse is true for large sample sizes. The magnitude of the differences is small, particularly when compared to the results for country-specific and pooled estimates. When compared to the observed country-specific difference in means, vector shrinkage using the weighted estimator for the between-country variance results in a 21–59 per cent reduction in average standard error, with larger reductions for smaller sample sizes.

Figures 3(a) and 3(b) show the corresponding results for effectiveness. These are very similar to those for costs, the main differences being improved coverage of the random effects estimator, now ranging from 72 to 78 per cent, and a minimum coverage probability of 94 per cent for the shrinkage estimators. Compared to the observed country-specific difference in means, vector shrinkage estimation with the weighted estimator for the between-country variance achieved a 25–45 per cent reduction in average standard error, once again with smaller sample sizes resulting in the largest reductions.

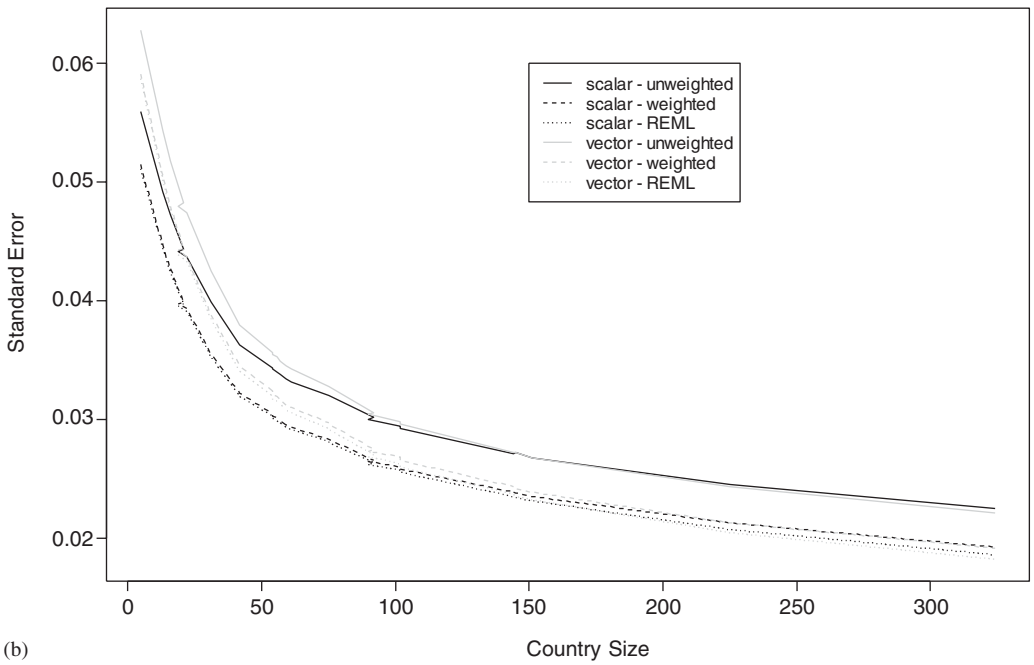
These findings are essentially unchanged when using a uniform rather than a normal distribution to generate the random effects. For costs, coverage probabilities fall to 92–94 per cent for the five largest countries (sample sizes 144–324) but otherwise all vector shrinkage methods result in coverage probabilities of at least 95 per cent. Scalar shrinkage estimates using the weighted or unweighted estimators for U have 94 per cent coverage for the next four largest countries (sample sizes 90–102), but are otherwise 95 per cent or higher. The random effects pooled estimator continues to have poor coverage at 51–59 per cent, depending on sample size and choice of estimator. The benefits of using a shrinkage estimator (vector, weighted estimator for U) over the observed country-specific means remain at a 21–59 per cent reduction in average standard error, with larger reductions for smaller sample sizes. For effectiveness, coverage probabilities are all 95 per cent or higher, with the exception of the two largest countries. Compared to the observed country-specific difference in means, vector shrinkage with the weighted estimator for U results in a 25–59 per cent reduction in average standard error.

5. DISCUSSION

The simulation results show that for the ASSENT data, shrinkage estimators result in adequate interval coverage and substantial reductions in standard errors compared to country-specific estimation. Coverage probabilities are related to country sample size, with excessive coverage in smaller countries and a slight shortfall in larger countries. At one level it is fortunate that coverage probability should be inversely related to sample size, since it is for the small to moderately sized country that the shrinkage estimator is most useful. If there is doubt about the coverage probability of the shrinkage estimator in a country with a large sample size, there will often be sufficient data for an informative country-specific analysis. However, the excessive coverage for small countries suggests that the standard errors are too large, and hence that further gains in precision should be possible. The expressions for the posterior variance of the true country means θ_i are derived in the equal variance case, where all countries recruit equal numbers of patients, and hence are an approximation in the more general (and more realistic) case of unequal sample sizes. Since the inequality in sample sizes leads to algebraically intractable posterior distributions, a natural approach would be to use Markov chain Monte Carlo (MCMC) methods to simulate from the posterior distributions for the θ_i.



(a)



(b)

Figure 3. (a) Coverage probabilities for effectiveness by number of patients per treatment arm; and (b) effectiveness average standard errors by number of patients per treatment arm.

This should lead to better coverage in countries with larger sample sizes and better precision in countries with small sample sizes. Since the simulations reported here have examined only one particular distribution of sample sizes, an MCMC approach might be preferable for problems whose country sample sizes are more disparate than those in the ASSENT trial.

Since the interval estimates described here are derived using the shrinkage estimator and the posterior variance based on placing vague hyperpriors on the unknowns τ and U, they can be viewed as Bayesian credible intervals [38]. The simulation results evaluate the correctness of the intervals neither as frequentist confidence intervals nor directly as Bayesian intervals. To evaluate them as the former, the 'true' country-specific means θ_i need to remain fixed from one simulated data set to the next. Doing so results in coverage probabilities that are excessive for those countries whose θ_i are close to the overall mean τ and far too small (as low as 70 per cent) for countries with means in the tails of the distribution, meaning that the intervals cannot be given a frequentist interpretation. In contrast, evaluating the intervals as Bayesian intervals would require the data to remain fixed while the 'true' means θ_i varied. If the formulae for the shrinkage estimators and their posterior variances were exactly correct and the posterior distributions of each θ_i were normal, the interval estimates would be exact Bayesian credible intervals. In this case, the coverage probabilities from the simulations would be providing information about calibration: that is, if the same vague hyperpriors for τ and U are used in a large number of estimation problems where the true distribution generating the parameters θ_i is $\mathrm{N}(\tau_0, U_0)$, with τ_0 and U_0 some known constants, do 95 per cent of the resulting 95 per cent credible intervals cover their unknowns [39]? Since the formulae for the posterior variance of the θ_i and the assumption of posterior normality are approximations, coverage probabilities which are different from 95 per cent could arise either from incorrect credible intervals (i.e. from the approximations themselves), or from incorrect calibration (i.e. the use of vague hyperpriors for τ and U when in the simulations these are held fixed). In practice, the first source of error can be overcome by replacing the approximate intervals with Bayesian credible intervals based on MCMC estimates of the posterior distributions. The problem of calibration can be avoided by using appropriate priors for the unknown hyperparameters, based on the particular estimation problem at hand. An appropriate prior selection strategy should result in correct calibration (see Reference [39]).

Empirical Bayes shrinkage estimation makes two sets of distributional assumptions. The first is that the observed differences in means follow a normal distribution with a country-specific mean and a country-specific variance. Most approaches to cost–effectiveness analysis of clinical trials data make this assumption and it has been shown to hold approximately even when patient-level observations are quite skewed [4, 40–42]. Its validity may be questionable for countries which contribute small numbers of patients; however, the variance associated with the treatment means for such countries is high, meaning that they have little influence on the shrinkage results.

The second assumption is that the true country-specific means are drawn from a normal distribution with mean and variance common to all countries. In some circumstances, structural differences between countries may imply that a common mean is not a realistic assumption, and it will be appropriate to differentiate between, for example, developing and developed countries. This can be done by incorporating country-level covariates into the mean of the second-level normal distribution. In a frequentist context, the assumption is that the (possibly adjusted) random effects for country are normally distributed, and is a stronger assumption whose validity cannot rely on the central limit theorem. However, Morris [29] showed that,

provided countries recruit equal numbers of patients, the shrinkage estimators derived using a normal prior are empirical Bayes minimax. Informally, this means that conclusions are conservative under departures from normality (see Reference [32] for further discussion of minimaxity). The simulation results reported here provide some evidence of robustness in the unequal variance case, although they are relevant only to a specific departure from normality, a specific selection of within-country variances and a particular between-country variance. When viewed as a Bayesian estimation problem, the assumption of a normal prior is made conditionally on the prior parameters τ and U. Since these are unknown, they are given vague hyperpriors, which in the scalar case are a Jeffrey's prior for τ and a Uniform$[0, \infty)$ for U. Integrating out over these gives priors for the θ_i which are parameter-free and no longer normal: for example in the scalar case, integrating first over U gives a prior for θ_i, conditionally on τ that is proportional to $|\theta_i - \tau|$. Thus the actual prior used for the θ_i is not normal, and is so diffuse that it will be dominated by the data in most estimation problems.

A Bayesian analysis would solve a few of the problems raised by this empirical Bayes approach: the credible intervals would have the right coverage and using a vector approach would likely lead to smaller standard errors than a scalar approach. What then is the value of the empirical Bayes estimators? Their robustness property is one answer. Whilst it is true that the Bayesian credible intervals will have the right coverage, this is the case only if the prior is universally accepted. Thus it may be helpful to include a sensitivity analysis, calculating estimators for a range of plausible priors. If the results are robust to changes in the prior, then both the empirical Bayes and the Bayesian estimator are robust to (reasonable) prior misspecification. If, however, the Bayesian results depend importantly on the choice of prior, then different researchers will have different Bayesian estimators. It is in this scenario that the empirical Bayes estimator is useful.

The six different shrinkage estimators considered here performed quite similarly to one another. There was a tendency for those based on the unweighted estimators for U to have both larger standard errors and larger coverages than the other estimators. REML and weighted estimators performed very similarly to one another, for both vector and scalar shrinkage. This would suggest favouring the weighted estimator for U, unless a regression approach is used, since the REML estimator needs to be evaluated iteratively. If just costs or just effectiveness are of interest, then a scalar estimator could be used, since the benefits of vector estimation, if any, appear to be minimal. For cost–effectiveness analyses vector shrinkage may yield a better estimate of the posterior covariance between costs and effects, although this was not evaluated empirically.

An important simulation result is that, in the presence of between-country heterogeneity, the random effects pooled estimator is a poor estimator of country-specific effects. Moreover, compared to the observed difference in country means, the shrinkage estimators achieved a 20–59 per cent reduction in average standard error. For Canada there was a 33 per cent reduction, and achieving this gain in precision using only Canadian data would have required more than twice as many patients.

ACKNOWLEDGEMENTS

This work was partially funded by a grant from Aventis. The authors wish to thank Ms Susan Tomlinson for her meticulous proof reading. Of course, all remaining errors are our own responsibility. B.J.O. was supported by a Senior Investigator Award from the Canadian Institutes of Health Research,

E.M.P. is supported through a studentship, fully or in part, by the Ontario Student Opportunity Trust Fund-Hospital for Sick Children Foundation Student Scholarship Programme, and A.R.W. is funded through the Discovery Grant Programme of the Natural Science and Engineering Research Council of Canada (Grant number 44868-03), however the opinions expressed are those of the authors and should not be attributed to any funding agency. The authors wish to thank the reviewers for the many helpful comments, which have led to a much improved manuscript.

REFERENCES

1. O'Brien BJ, Drummond MF, Labelle RJ, Willan AR. In search of power and significance: issues in the design and analysis of stochastic cost–effectiveness studies in heath care. *Medical Care* 1994; **32**:150–163.
2. Mullahy J, Manning W. Statistical issues of cost–effectiveness analysis. In *Valuing Health Care*, Sloan F (ed.). Cambridge University Press: Cambridge, 1994; 149–184.
3. Wakker P, Klaassen MP. Confidence intervals for cost/effectiveness ratios. *Health Economics* 1995; **4**:373–381.
4. Willan AR, O'Brien BJ. Confidence intervals for cost–effectiveness ratios: an application of Fieller's theorem. *Health Economics* 1996; **5**:297–305.
5. Chaudhary MA, Stearns SC. Confidence intervals for cost–effectiveness ratios: an example from a randomized trial. *Statistics in Medicine* 1996; **15**:1447–1458.
6. Mullahy J. What you don't know can't hurt you? Statistical issues and standards for medical technology evaluation. *Medical Care* 1996; **34**(12 Suppl.):DS124–DS135.
7. Manning WG, Fryback DG, Weinstein MC. Reflecting uncertainty in cost effectiveness analysis. In *Cost Effectiveness in Health and Medicine*, Gold MR, Siegel JE, Russell LB, Weinstein MC (eds). Oxford University Press: New York, 1996.
8. Briggs AH, Wonderling DE, Mooney CZ. Pulling cost–effectiveness analysis up by its bootstraps; a non-parametric approach to confidence interval estimation. *Health Economics* 1997; **6**:327–340.
9. Polsky D, Glick HA, Willke R, Schulman K. Confidence intervals for cost–effectiveness ratios: a comparison of four methods. *Health Economics* 1997; **6**:243–252.
10. Willan AR, O'Brien BJ. Sample size and power issues in estimating incremental cost–effectiveness ratios from clinical trials data. *Health Economics* 1999; **8**:203–211.
11. Phelps CE, Mushlin AI. On the (near) equivalence of cost–effectiveness and cost–benefit analysis. *International Journal of Technology Assessment in Health Care* 1991; **7**:12–21.
12. Ament A, Baltussen R. The interpretation of results of economic evaluation: explicating the value of health. *Health Economics* 1997; **6**:625–635.
13. Stinnett AA, Mallahy J. Net health benefits: a new framework for the analysis of uncertainty in cost–effectiveness analysis. *Medical Decision Making* 1998; **18**(Suppl.):S68–S80.
14. Tambour M, Zethraeus N, Johannesson M. A note on confidence intervals in cost-effectiveness analysis. *International Journal of Technology Assessment* 1998; **14**:467–471.
15. van Hout BA, Al MJ, Gordon GS, Rutten FFH. Costs, effects and C/E ratios alongside a clinical trial. *Health Economics* 1994; **3**:309–319.
16. Briggs A, Fenn P. Confidence intervals or surfaces? Uncertainty on the cost–effectiveness plane. *Health Economics* 1998; **7**:723–740.
17. Briggs AH. A Bayesian approach to stochastic cost–effectiveness analysis. *Health Economics* 1999; **8**:257–261.
18. Lothgren M, Zethraeus N. Definition, interpretation and calculation of cost–effectiveness acceptability curves. *Health Economics* 2000; **9**:623–630.
19. Heitjan DF. Fieller's method and net health benefit. *Health Economics* 2000; **9**:327–335.
20. Willan AR. Analysis, sample size and power for estimating incremental net health benefit from clinical trial data. *Controlled Clinical Trials* 2001; **22**:228–237.
21. Willan AR, Lin DY. Incremental net benefit in randomized clinical trials. *Statistics in Medicine* 2001; **20**: 1563–1574.
22. Willan AR, Chen EB, Cook RJ, Lin DY. Incremental net benefit in clinical trials with quality-adjusted survival. *Statistics in Medicine* 2003; **22**:353–362.
23. Willan AR, Lin DY, Cook RJ, Chen EB. Using inverse-weighting in cost–effectiveness analysis with censored data. *Statistical Methods in Medical Research* 2002; **11**:539–551.
24. Cook JR, Drummond M, Glick H, Heyse JF. Assessing the appropriateness of combining economic data from multinational clinical trials. *Statistics in Medicine* 2003; **22**:1955–1976.
25. Koopmanschap MA, Touw KCR, Rutten FFH. Analysis of cost and cost–effectiveness in multinational trials. *Health Policy* 2001; **58**:175–186.
26. Wilke RJ, Glick HA, Polsky D, Schulman K. Estimating country-specific cost–effectiveness from multinational clinical trials. *Health Economics* 1998; **7**:481–493.

27. Manca A, Rice N, Sculpher MJ, Briggs AH. Assessing generalisability by location in trial-based cost–effectiveness analysis: the use of multilevel models. *Health Economics* 2004 (www.interscience.wiley.com DOI: 10 1002/hec.914).
28. James W, Stein C. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. University of California Press: Berkeley, 1961; 361–370.
29. Morris C. Interval estimation for empirical Bayes generalizations of Stein's estimator. *Proceedings of the Twenty-Second Conference on the Design of Experiments in Arym Research Development and Testing*, *ARO Report 77-2*, 1977.
30. Willan AR, Pinto EM, O'Brien BJ, Kaul P, Goeree R, Lynd L, Armstrong PW. Country-specific cost comparisons from multinational clinical trials using empirical Bayesian shrinkage estimation: the Canadian ASSENT-3 Economic Analysis. *Health Economics*, (to appear).
31. Efron B, Morris C. Empirical Bayes on vector observations: an extension of Stein's method. *Biometrika* 1972; **59**:335–347.
32. Morris C. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* 1983; **78**:47–55.
33. Dersimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
34. The ASSENT-3 Investigators. Efficacy and safety of tenecteplase in combination with enoxaparin, abciximab, or unfractionated heparin: the ASSENT-3 randomised trial in acute myocardial infarction. *Lancet* 2001; **358**: 605–613.
35. Wallentin L, Lindahl B, Armstrong PW, Granger CB, Van de Werf F, ASSENT 3 Investigators. Influence of enoxaparin and early revascularization on death or myocardial infarction after thrombolysis in ST elevation myocardial infarction. *European Heart Journal* 2002; **23**(Suppl.):639.
36. Ontario Case Cost Project. *Ontario Case Cost Project* (*OCCP*): *Ontario Guide to Case Costing*, *Version* 1.1. Ontario Case Cost Project: Ottawa, September, 1995.
37. Ontario Ministry of Health. *Schedule of Benefits*: *Physician Services Under the Health Insurance Act*, February 1, 1998. Queen's Printer: Toronto, Ontario, 1999.
38. Berger J. Comment on parameteric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* 1983; **78**:55–57.
39. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* 1984; **12**:1151–1172.
40. Thompson SG, Barber JA. How should cost data in pragmatic clinical trials be analysed? *British Medical Journal* 2000; **320**:1197–1200.
41. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. *Annual Review of Public Health* 2002; **23**:151–169.
42. Willan AR, Briggs AH, Hoch JS. Regression methods for covariate adjustment and subgroup analysis for non-censored cost–effectiveness data. *Health Economics* 2004; **13**:461–475.

