An overview of methods and empirical comparison of aggregate data and individual patient data results for investigating heterogeneity in meta-analysis of time-to-event outcomes

Catrin Tudur Smith MSc,[1] Paula R. Williamson PhD[2] and Anthony G. Marson MD[3]

[1]Research Associate, Centre for Medical Statistics and Health Evaluation, University of Liverpool, Liverpool, UK
[2]Director and Reader, Centre for Medical Statistics and Health Evaluation, University of Liverpool, Liverpool, UK
[3]Senior Lecturer in Neurology, Department of Neurological Sciences, University of Liverpool, Liverpool, UK

Correspondence
Catrin Tudur Smith
Centre for Medical Statistics and Health Evaluation
Shelley's Cottage
University of Liverpool
Liverpool L69 3GS
UK
E-mail: cat1@liv.ac.uk

Keywords: aggregate data (AD), heterogeneity, individual patient data (IPD), meta-regression, time-to-event outcomes

Accepted for publication:
1 September 2004

Abstract

Combining the results of individual studies using meta-analysis may be undertaken using either aggregate data (AD) or individual patient data (IPD). In any meta-analysis it is important to consider statistical heterogeneity between studies. Potential sources of heterogeneity can be explored using regression models with either AD or IPD. An overview of approaches and empirical assessment of how the results and conclusions differ from these analyses is undertaken using a meta-analysis of five randomized controlled trials comparing two antiepileptic drugs with time-to-event outcomes. Alternative meta-regression models using AD are compared to stratified Cox regression models using IPD. Age as a potential cause of heterogeneity is detected by both AD and IPD regression models. Time from first ever seizure to randomization is only identified by some AD models. A more thorough explanation of heterogeneity is obtained from the model using IPD but further empirical evidence comparing IPD and AD results are needed.

Introduction

In meta-analysis, the collection and use of individual patient data (IPD) has been described as the 'gold-standard' approach (Chalmers 1993). The process of collecting IPD can be a resource intensive undertaking, but is valuable when time-to-event outcomes are considered (Stewart & Clarke 1995). The more common alternative approach to meta-analysis uses aggregate data (AD) presented in individual study reports or made available by the trialists. For time-to-event outcomes, methods are available to estimate the relative treatment effect (log hazard ratio) and its variance from AD (Parmar *et al*. 1998; Williamson *et al*. 2002b). As an example, these statistics can be estimated from the *P*-value of the logrank test and number of events. These methods are useful if IPD are not available, or preliminary analyses are to be undertaken to assess whether collecting IPD is worthwhile. However, it is frequently the case that sufficient information is not presented in the study report (Altman *et al*. 1995) and consequently some of these methods cannot be used. Furthermore, one of the methods based on survival curves may not always provide reliable estimates especially when the event rate is low (Tudur *et al*. 2001). Williamson *et al*. (2000) undertook a review of 'reviews which have included a comparison of the main treatment effect results from an IPD and AD meta-analysis'. Although the comparisons examined and methods used within each review varied, a general comparison between the 'gold standard' IPD and the literature based AD meta-analysis could be made. The statistical significance and estimate of the treatment effect obtained from the two approaches varied but the direction of differences was not consistent across reviews. Further empirical evidence with more specific comparisons were suggested in order to establish whether the extra investment needed for the IPD approach, over and above AD, is worthwhile.

In any meta-analysis, or any study comparing meta-analytic methods, it is important to consider statistical heterogeneity; variation in the true treatment effect between studies. Differing design features, methodological quality, variability in clinical procedures and patient characteristics are all factors that can contribute to such variability and should be investigated (Thompson 1994). Regression modelling is a popular approach for investigating heterogeneity between studies and to assess the effect of important characteristics. If IPD are available, patient level characteristics are included in the model and relationships with treatment and the impact on heterogeneity can be investigated. With AD, study-level characteristics are included in the model and relationships with the study-level response, corresponding to the relative treatment effect in a trial, are examined. In the meta-analysis literature such regression approaches are called meta-regression. Alternative model structures and approaches for parameter estimation are available and differ depending on whether IPD or AD are used in the model. There is a need therefore to compare the results and conclusions obtained from meta-regression using these two data approaches.

Individual patient data are available from a systematic review comparing two drugs, carbamazepine (CBZ) and sodium valproate (VPS), for the treatment of epilepsy (Marson *et al.* 2000) in which statistical heterogeneity between trials was evident for one outcome. The data available from this review are used to illustrate and compare potential explanations of heterogeneity when using regression models with IPD or AD. In this example, a pragmatic comparison of IPD and AD extracted from study reports was not possible as appropriate AD could not be extracted. To allow a comparison between IPD and AD approaches, the IPD were therefore used to generate AD that may typically be presented in trial reports.

In the following sections, we summarize regression models for exploring sources of heterogeneity with aggregate or individual patient time-to-event data and apply each approach to the example from epilepsy. Results and conclusions drawn are compared to provide further insight and empirical evidence of the comparison between AD and IPD.

Example from epilepsy

Time-to-event outcomes were examined in an IPD systematic review of randomized controlled trials (RCTs) comparing CBZ and VPS (Marson *et al.* 2000), both commonly used drugs for the treatment of epilepsy. Information for eight pre-specified clinically important patient characteristics [age at randomization, sex, seizure type (generalized or partial), number of seizures prior to randomization, time from first seizure to randomization, results of electroencephalograph (EEG) scan, results of computed tomography (CT) scan, neurological signs at clinical examination] were requested from each trial. However, as a result of the high proportion of missing information for the latter three variables their impact are not examined in subsequent models. Natural logarithms were taken if the distribution of a continuous variable appeared skewed (number of seizures in 6 months pre-randomization, time from first seizure to randomization).

A stratified logrank analysis was originally undertaken to investigate the main effect of drug and the test for homogeneity suggested evidence of statistically significant heterogeneity for the outcome 'time to 12 month remission' ($\chi^2_4 = 11.75$, $P = 0.02$). In the original review, an explanation for heterogeneity was sought by examining the test for an interaction between drug and patient factors (age at randomization, sex, seizure type, number of seizures prior to randomization and time from first ever seizure to randomization) using univariate logrank analyses. A Cox regression model with trial indicator variables was subsequently used (Williamson *et al.* 2002a) and a possible explanation of heterogeneity obtained. These results motivated the question as to whether the same explanation of heterogeneity and clinical interpretations would have been obtained had the IPD not been available.

The AD generated from IPD for the outcome 'time to 12 month remission' and the five patient factors of interest, are summarized in Table 1.

Table 1 CBZ–VPS example: aggregate data generated for each trial

Trial	*Remission events/total number*	*Log hazard ratio* (SE)*	*Mean age*	*Proportion female*	*Proportion partial epilepsy*	*Mean (log(number of seizures))*	*Mean (log(time from first ever seizure))*
1. Heller *et al.* 1995	81/122	–0.086 (0.223) *n* = 122	30.65 *n* = 119	0.52 *n* = 122	0.40 *n* = 122	1.27 *n* = 119	0.81 *n* = 119
2. de Silva *et al.* 1996	88/103	0.223 (0.214) *n* = 103	10.19 *n* = 103	0.53 *n* = 103	0.52 *n* = 103	1.79 *n* = 103	0.49 *n* = 103
3. Richens *et al.* 1994	224/288	–0.371 (0.134) *n* = 288	33.32 *n* = 286	0.50 *n* = 288	0.49 *n* = 288	1.67 *n* = 288	N/A
4. Verity *et al.* 1995	183/246	0.158 (0.148) *n* = 246	10.09 *n* = 244	0.53 *n* = 246	0.42 *n* = 246	1.56 *n* = 245	0.08 *n* = 226
5. Mattson *et al.* 1992	191/466	–0.312 (0.145) *n* = 466	47.21 *n* = 466	0.08 *n* = 466	1.00 *n* = 466	3.01 *n* = 431	1.46 *n* = 450

*The log hazard ratio for VPS compared to CBZ calculated using individual Cox regression models with Efron method for handling ties for each trial (a positive log hazard ratio indicates a clinical advantage for VPS).
N/A, not available since covariate not measured in this trial; SE, standard error; CBZ, carbamazepine; VPS, sodium valproate.

Regression models for investigating heterogeneity

Regression models can be used to investigate associations between patient, or study characteristics and the outcome, or summary effect measure of interest, depending on whether IPD or AD are used. These models are used to identify potential sources, and attempt to explain or reduce statistical heterogeneity by including important covariates in the model. The following sections provide a brief summary of suitable regression models for investigations involving time-to-event outcomes.

Regression using individual patient data

A Cox proportional hazards regression model stratified by trial given by

$$\lambda_{ij}(t) = \lambda_{oj}(t)\exp(\beta_1 x_{1ij}) \qquad (1)$$

for the *i*th individual in the *j*th trial with treatment indicator variable $x_{1,}$ can be used to undertake meta-analysis and further assess the effect of patient-level characteristics by incorporating covariates into the linear predictor of the model. Comparison of model (1) with a stratified Cox model that includes study specific treatment effects can be used to assess the evidence for heterogeneity. An alternative to the Cox model stratified by trial would be to include trial indicator variables in a standard Cox regression model but this is not investigated here as it makes the rather restrictive assumption that the hazards are proportional overall compared to the less restrictive assumption (that the hazards are proportional within each trial rather than overall) made by the stratified model.

In model (1) the relative treatment effect (log hazard ratio), denoted by parameter β_1, is assumed to be identical across trials (fixed effects model). An alternative approach allowing the relative treatment effect to vary across trials can be achieved by including treatment as a random effect,

$$\lambda_{ij}(t) = \lambda_{oj}(t)\exp(\beta_{1j} x_{1ij}) \qquad (2)$$

$$\beta_{1j} = \beta_1 + \mathrm{b}_{1j}$$

where the random quantities b_{1j} are assumed to follow a normal distribution with mean zero and variance τ^2 which is a measure of the between trial variability in treatment effect (statistical heterogeneity). Further details and examination of IPD models including random trial effects are given by Tudur Smith *et al.* (2005).

Meta-regression using aggregate data

Investigating heterogeneity and the influence of prognostic factors in a regression framework is commonly undertaken using AD and the phrase meta-regression used to describe these models. A brief outline of the methods considered in our comparison is given below but a discussion of assumptions and further details of the models are described in depth by Thompson & Sharp (1999). As only five trials are available in the antiepileptic drug example multivariate models using AD are not investigated.

The first model using AD assumes the observed log hazard ratio in each trial ($\log HR_j$) is independently normally distributed such that

$$\log HR_j \sim N(\alpha + \beta x_j, v_j) \tag{3}$$

where v_j is the variance of the log hazard ratio in the jth trial. In order to account for the assumption that the variance of the log hazard ratio from each trial is not equal, a weighted regression with weights defined by the reciprocal of the variance is used. This model is fitted using standard statistical software for weighted regression with the standard errors (SEs) corrected through division with the square root of the mean squared error (MSE) (Thompson & Sharp 1999).

This first meta-regression model (3) corresponds to a fixed effect model as between trial variance is not accounted for. This model can be extended to incorporate heterogeneity that is left unexplained by the fitted covariates through inclusion of an additive between study variance component τ^2. This 'random effects' meta-regression model may be written as follows:

$$\log HR_j \sim N(\alpha + \beta x_j, v_j + \tau^2) \tag{4}$$

An explicit estimate of τ^2 is needed as the weights used in the regression are given by the inverse of the sum of the within and between study variance components. The Moment Estimator (MM), Maximum Likelihood estimate (ML), Restricted Maximum Likelihood estimate (REML) and Empirical Bayes estimate (EB) methods of estimating τ^2 have been proposed and are described in detail by Thompson & Sharp (1999).

Results

Regression using individual patient data

Model (1) was fitted using the *coxph* function within the R computer program. A SAS IML program has been developed to enable model (2) to be fitted. Further details are available from the first author and described in more detail elsewhere (Tudur Smith *et al.* 2005). The evidence against the assumption of homogeneity, assessed by comparing –2log(likelihood) of model (1) to the corresponding model including study specific treatment effects, suggested evidence for statistical heterogeneity ($\chi^2_4 = 11.30$, $P = 0.02$). The between trial variability in log hazard ratio (τ^2) is estimated to be 0.048 with SE 0.055 using the random treatment effect approach (model 2). Both models suggest a non-significant benefit in favour of CBZ although the parameter estimate of β_1 and its SE obtained from both models differ somewhat; –0.132 (0.073) from model (1) and –0.098 (0.125) from model (2). The larger SE from model (2) is to be expected as allowance has also been made for heterogeneity in this random effects approach.

The parameter estimates obtained from univariate stratified Cox regression models (using model 1) suggested a significant main effect of seizure type and log(number of seizures in 6 months pre-randomization). To enable a comparison to be made with meta-regression models using AD, treatment by covariate interaction terms were added to each individual model including main effects of treatment and the covariate of interest. The only significant interaction, as identified from model (1), appears to be between treatment and age (Table 2). Older people on CBZ have a better clinical outcome than younger people on CBZ, whereas VPS has a similar effectiveness across the age range.

To arrive at an explanation for heterogeneity the following model selection strategy was adopted: interaction with drug terms were fitted in the model which included all covariates significant on univariate analyses and non-significant terms were subsequently dropped from this model. The 'final' model (Table 3) included, treatment, seizure type, log(number of seizures), age and an interaction between age and treatment. Having considered these patient-level covariates in the fixed treatment effect model (1), the test for heterogeneity is no longer statistically significant ($P = 0.44$) and suggests that the included covariates provide a possible explanation for the heterogeneity in treatment effect across trials. Furthermore, in support of these results, the estimate of τ^2 from the random effects model (2) decreases after allowing for the same covariates.

Table 2 Parameter estimates (SE) from univariate stratified Cox models with main effect of treatment, covariate and corresponding interaction term using IPD

Model	*Covariate*	*Treat (VPS) (SE)*	*Covariate (SE)*	*Treat × covariate (SE)*	*Change, d.f., P-value*[†]	τ^2 *(SE)*
(1)	Null	−0.132 (0.073)	–	–		
	Age	0.210 (0.129)	0.004 (0.003)	−0.012 (0.004)	9.99, 1, $P = 0.002$	
	Sex (female)	−0.086 (0.095)	−0.016 (0.081)	−0.110 (0.148)	0.55, 1, $P = 0.46$	
	Epilepsy type (partial)	−0.043 (0.113)	−0.351 (0.085)	−0.150 (0.147)	1.04, 1, $P = 0.31$	
	Log(number of seizures)	−0.048 (0.122)	−0.214 (0.030)	−0.028 (0.057)	0.25, 1, $P = 0.62$	
	Log(time from first ever seizure)	0.037 (0.107)	−0.057 (0.050)	−0.081 (0.086)	0.89, 1, $P = 0.35$	
(2)	Null	−0.098 (0.125)	–	–		0.0484 (0.055)
	Age	0.210 (0.129)	0.004 (0.003)	−0.012 (0.004)		0
	Sex (female)	−0.006 (0.152)	−0.030 (0.081)	−0.206 (0.158)		0.064 (0.067)
	Epilepsy type (partial)	−0.049 (0.151)	−0.347 (0.085)	−0.088 (0.160)		0.043 (0.052)
	Log(number of seizures)	−0.035 (0.147)	−0.213 (0.030)	−0.019 (0.058)		0.031 (0.043)
	Log(time from first ever seizure)	0.005 (0.150)	−0.057 (0.051)	−0.013 (0.095)		0.040 (0.063)

[†] The *P*-values are obtained from comparing the change in −2logL (fixed effect model 1) with chi-square critical values.
IPD, individual patient data; SE, standard error; VPS, sodium valproate.

Table 3 Parameter estimates (SEs), hazard ratios and their 95% confidence interval (CI) for 'final' stratified Cox regression models using IPD (based on 1183 individuals)

	Model (1)		*Model (2)*	
Covariate	*Coefficient (SE)*	*Hazard ratio (95% CI)*	*Coefficient (SE)*	*Hazard ratio (95% CI)*
Treat (VPS)	0.162 (0.129)	1.18 (0.91, 1.51)	0.163 (0.139)	1.18 (0.90, 1.55)
Seizure type (partial)	−0.186 (0.088)	0.83 (0.70, 0.99)	−0.185 (0.088)	0.83 (0.70, 0.99)
Log(number of seizures)	−0.192 (0.031)	0.83 (0.78, 0.88)	−0.192 (0.031)	0.83 (0.78, 0.88)
Age at randomization	0.005 (0.003)	1.01 (1.00, 1.01)	0.005 (0.003)	1.01 (1.00, 1.01)
Age by drug interaction	−0.009 (0.004)	0.99 (0.98, 1.00)	−0.009 (0.004)	0.99 (0.98, 1.00)
τ^2 (SE)	NA		0.006 (0.027)	

IPD, individual patient data; NA, not available; SE, standard error; VPS, sodium valproate.

As the variables age at randomization and number of seizures before randomization were not recorded for 42/1225 individuals, the parameter estimates displayed in Table 3 are based on a subset of 1183 individuals. To enable a comparison of parameters before and after inclusion of these covariates, each corresponding model including only a treatment variable were fitted to the same subset of 1183 individuals (Table 4).

On comparison of parameter estimates from Tables 3 and 4, the percentage change in τ^2 (from 0.043 to 0.006) suggests that inclusion of age, epilepsy type, log(number of seizures), and an age by treatment interaction has explained 86% of the heterogeneity. From a clinical perspective, the age by treatment interaction suggests that older patients taking CBZ are more likely to experience a period of 12 month remission from seizures, hence a better clinical outcome, whilst younger patients fare better on VPS. In general, this change in direction of effect occurs at around the age of 18. Further details regarding the clinical implication of these results are given by Marson *et al*. (2000) and Williamson *et al*. (2002a).

Table 4 Parameter estimates (SEs), hazard ratios and their 95% confidence interval (CI) for stratified Cox regression models without covariates using IPD (based on 1183 individuals)

	Model (1)		*Model (2)*	
Covariate	*Coefficient (SE)*	*Hazard ratio (95% CI)*	*Coefficient (SE)*	*Hazard ratio (95% CI)*
Treat (VPS)	–0.112 (0.074)	0.89 (0.77, 1.03)	–0.081 (0.120)	0.92 (0.73, 1.17)
τ^2 (SE)	NA		0.043 (0.051)	

IPD, individual patient data; NA, not available; SE, standard error; VPS, sodium valproate.

Meta-regression using aggregate data

Models (3) and (4) were fitted using STATA 6.0. The parameter estimates and corresponding SEs obtained from fitting univariate models are summarized in Table 5.

As residual heterogeneity is not accounted for in model (3) the SEs of each regression coefficient are smaller (or equivalent in some cases) compared to model (4). There is generally reasonable agreement between the estimated regression coefficients from alternative models using AD.

All models provide strong evidence for a significant effect of mean age and mean (log(time from first ever seizure)) with identical regression coefficients and SEs for these variables across all models. The results suggest that VPS is more effective for younger patients, whilst CBZ is more effective for older patients with the change in direction of effect occurring at around age 19. VPS appears more effective for shorter intervals between first seizure and randomization (less than approximately 2 years) and CBZ more effective for larger intervals (greater than approximately 2 years). For model (4), the between trial variability parameter τ^2 is estimated to be zero for all estimation procedures when either of these variables are included suggesting that all of the heterogeneity may be explained by considering these trial level covariates. This is further supported by noting that the parameter estimates and SEs for random effects models (4) are precisely equal to those of the fixed effect model (3) after including these covariates.

Evidence to suggest a relationship between treatment effect and any other covariate examined in this investigation is much weaker, with non-zero estimates of τ^2 indicating that some residual variability remains unexplained by the effect of each aggregate level covariate. For model (4) comparing the change in τ^2 with corresponding null models as a measure of the proportion of variation explained gives quite different values depending on the estimation approach. However, as estimation of τ^2 is poor when the number of included trials is small, as in this case, the reliability of these results is questionable and care is required for interpretation. In this example with five trials, estimates of τ^2 for all covariates are smaller using ML compared with MM, REML and EB approaches. Thompson & Sharp (1999) suggest that the maximum likelihood approaches are preferable but because of the downward bias of the estimate of τ^2 using ML, they suggest that an REML estimate will be most appropriate in practice.

Comparing individual patient data and aggregate data results

Parameter estimates obtained from models using IPD (Table 2) or AD (Table 5) may be compared to provide an empirical evaluation of meta-regression analyses using both types of data. In the CBZ–VPS example, AD were generated from IPD and the comparison is therefore slightly artificial and reflects a comparison of methods rather than results that might be seen in reality but could be affected by other reporting factors.

The first point to note is the AD null model result assuming fixed treatment effect (model 3, Table 5) gives exactly the same parameter estimate and SE [–0.132 (0.073)] as the stratified Cox model using IPD (model 1, Table 2). This agreement occurs as expected because the AD estimates of log hazard ratio and SE (Table 1) have been generated from

Table 5 Parameter estimates (SE) from univariate meta-regression models using AD

Covariate		*α* (SE)*	*β (SE)*	*Z,* P*-value*	*τ²*
Model (3)	Null	–0.132 (0.073)	–	–	–
	Mean age	0.290 (0.158)	–0.015 (0.005)	–3.0, $P = 0.003$	–
	Proportion female	–0.385 (0.172)	0.621 (0.382)	1.63, $P = 0.104$	–
	Proportion partial	0.154 (0.196)	–0.481 (0.306)	–1.57, $P = 0.116$	–
	Mean (log(number of seizures))	0.192 (0.237)	–0.166 (0.116)	–1.43, $P = 0.152$	–
	Mean (log(time from first ever seizure))	0.237 (0.139)	–0.365 (0.147)	–2.48, $P = 0.013$	–
Model (4) MM	Null	–0.098 (0.126)	–	–	0.05
	Mean age	0.290 (0.158)	–0.015 (0.005)	–3.01, $P = 0.003$	0
	Proportion female	–0.392 (0.329)	0.700 (0.717)	0.98, $P = 0.329$	0.0563
	Proportion partial	0.199 (0.359)	–0.512 (0.577)	–0.89, $P = 0.376$	0.0578
	Mean (log(number of seizures))	0.228 (0.438)	–0.171 (0.220)	–0.78, $P = 0.437$	0.0612
	Mean (log(time from first ever seizure))	0.237 (0.139)	–0.365 (0.147)	–2.47, $P = 0.013$	0
Model (4) ML	Null	–0.103 (0.112)	–	–	0.034
	Mean age	0.290 (0.158)	–0.015 (0.005)	–3.01, $P = 0.003$	0
	Proportion female	–0.389 (0.243)	0.673 (0.533)	1.26, $P = 0.207$	0.0210
	Proportion partial	0.185 (0.269)	–0.504 (0.428)	–1.18, $P = 0.239$	0.0213
	Mean (log(number of seizures))	0.220 (0.330)	–0.171 (0.164)	–1.04, $P = 0.297$	0.0232
	Mean (log(time from first ever seizure))	0.237 (0.139)	–0.365 (0.147)	–2.47, $P = 0.013$	0
Model (4) REML	Null	–0.099 (0.124)	–	–	0.048
	Mean age	0.290 (0.158)	–0.015 (0.005)	–3.01, $P = 0.003$	0
	Proportion female	–0.391 (0.310)	0.696 (0.676)	1.03, $P = 0.303$	0.0475
	Proportion partial	0.198 (0.341)	–0.511 (0.547)	–0.93, $P = 0.351$	0.0497
	Mean (log(number of seizures))	0.227 (0.418)	–0.171 (0.209)	–0.82, $P = 0.415$	0.0534
	Mean (log(time from first ever seizure))	0.237 (0.139)	–0.365 (0.147)	–2.47, $P = 0.013$	0
Model (4) EB	Null	–0.099 (0.122)	–	–	0.045
	Mean age	0.290 (0.158)	–0.015 (0.005)	–3.01, $P = 0.003$	0
	Proportion female	–0.391 (0.299)	0.693 (0.652)	1.06, $P = 0.288$	0.0427
	Proportion partial	0.197 (0.334)	–0.510 (0.536)	–0.95, $P = 0.341$	0.0466
	Mean (log(number of seizures))	0.227 (0.411)	–0.171 (0.206)	–0.83, $P = 0.407$	0.0508
	Mean (log(time from first ever seizure))	0.237 (0.139)	–0.365 (0.147)	–2.47, $P = 0.013$	0

MM, Moment Estimator; ML, Maximum Likelihood; REML, Restricted Maximum Likelihood; EB, Empirical Bayes; AD, aggregate data; HR, hazard ratio; SE, standard error; CBZ, carbamazepine; VPS, sodium valproate.
*Comparison is VPS to CBZ therefore VPS is better if HR > 1.

IPD using a separate Cox model for each trial, but this would be unlikely to occur if AD were extracted from trial reports. The fixed effect AD meta-regression model (3) without covariates is equivalent to a simple inverse variance (IV) weighted average of trial level estimates. Assuming a fixed treatment effect, the IV weighted average of Cox model estimates can give very similar pooled results to those of the stratified Cox model under certain conditions. Further details are given by Tudur Smith (2004).

For models that assume random effects without considering the effect of covariates, parameter estimates and SEs for meta-regression analyses using AD (model 4, Table 5) are similar to the random effects model based on IPD (model 2, Table 2) which uses a REML approach for estimating τ^2.

The only treatment by covariate interaction identified as statistically significant by the stratified Cox model using IPD appears to be between treatment and age (Table 2). Very similar numerical results and clinical conclusions are drawn from the AD models (Table 5) after considering this particular covariate with both approaches estimating τ^2 to be equal to zero.

Based on the IPD Cox model results (Table 2), there is insufficient evidence to suggest that any further interactions exist between treatment and each of the covariates under consideration. Estimates of τ^2 that are close to that of the model without any covariate effects (null model in Table 2) suggest that inclusion of these other variables do not provide a sufficiently adequate explanation for heterogeneity. The AD model results (Table 5) agree to some extent in terms of the statistical significance for three covariates: proportion female, proportion partial epilepsy, and mean (log(number of seizures)). However, for mean (log(time from first ever seizure)), the AD models suggest evidence of a relationship ($P = 0.013$) with an estimate for τ^2 equal to zero which might suggest that this aggregate level variable can provide an explanation for statistical heterogeneity. As the safer IPD approaches failed to detect an overall within study relationship, the AD result is likely to be spurious and may be a result of multiple testing. This highlights the potential for misinterpretation that can arise when several associations with averages across a small number of trials are examined. For each of these variables there is often considerable disagreement in numerical results between IPD and AD approaches.

Assuming the AD generated from IPD had been available for each trial, one could have reached the conclusion that statistical heterogeneity could be explained by either the effect of age or time from first ever seizure. These two aggregate variables are highly correlated with longer average intervals from first ever seizure observed in trials with a greater average age. As data are available for a maximum of five trials, models including more than one covariate were not examined.

The availability of IPD allowed a thorough investigation into the main effects of each covariate (Table 3) which was not possible using meta-regression of AD. However, for the full stratified Cox model with random treatment effects (model 2, Table 3, based on 750 events and 1183 individuals as a result of missing covariate values) there is a small amount of residual heterogeneity [$\tau^2 = 0.006$ (0.027)] with 86% of the heterogeneity explained by including the main effects of age, epilepsy type, log(number of seizures) and an interaction between treatment and age. Let this model be referred to as model (A). The IPD model (2) that includes the main effect of age and an interaction with treatment term (Table 2, based on 764 events and 1218 individuals) suggests that 100% of the heterogeneity can be explained by these variables alone. Let this model be referred to as model (B). Because of a small amount of missing covariate data these two alternative models (A) and (B) are based on different subsets of the original data for 1225 individuals which makes a comparison of models difficult. As a sensitivity analysis, the variables treatment, age and their interaction term were fitted to the same subset of IPD for 1183 individuals used in model (A). The results based on this subset of data (Table 6) are not substantially different to the original (Table 2) but do suggest that a small amount

Table 6 Sensitivity analysis: parameter estimates (SE) from stratified Cox proportional hazards models with main effect of treatment, covariate and corresponding interaction term using IPD using subset of data for 1183 individuals

Model	*Covariate*	*Treat (VPS) (SE)*	*Covariate (SE)*	*Treat* covariate (SE)*	*Change, d.f.,* P-*value*	τ^2 *(SE)*
(1)	Null	–0.112	–	–		
		(0.074)				
	Age	0.206	0.004	–0.011	8.79, 1, $P = 0.003$	
		(0.130)	(0.003)	(0.004)		
(2)	Null	–0.081	–	–		0.0428
		(0.120)				(0.051)
	Age	0.204	0.004	–0.011		0.003
		(0.134)	(0.003)	(0.004)		(0.024)

IPD, individual patient data; SE, standard error; VPS, sodium valproate.

of residual heterogeneity remains unexplained [$\tau^2 = 0.003$ (0.024)] by these variables using these data. In summary, the age by treatment interaction appears to explain the heterogeneity across trials but the variables epilepsy type and log(number of seizures) are also clinically important.

Discussion

Meta-analyses are frequently undertaken using AD as, in many situations, IPD are simply not available. If time-to-event outcomes or particular subgroups are of interest, or a thorough investigation into potential sources of heterogeneity is required, IPD can be extremely valuable.

In this paper a brief overview is given of some existing meta-regression approaches for exploring sources of heterogeneity using aggregate time-to-event data. Corresponding analyses using IPD can be undertaken using the Cox regression model stratified by trial with either fixed or random treatment effects. These alternative approaches have been applied and compared to each other using a meta-analysis example evaluating two alternative treatments for epilepsy.

The example from epilepsy provided the original motivation to investigate and apply alternative models to undertake a meta-analysis and explore heterogeneity using individual patient failure time data. In some situations where trials agree in outcome definition and the reporting of suitable data, aggregate approaches are likely to be less resource intensive but potentially more restricted. A pragmatic comparison of results using IPD vs. results using extracted AD was not possible for this example as sufficient data were unavailable directly from trial reports. Such a limitation commonly arises in meta-analysis and often prevents any reasonable investigation into sources of heterogeneity. As the AD used for comparison were constructed from the IPD, the AD results represent the 'best possible' results obtainable using this data type. One advantage of having IPD is the ability to examine main effects of covariates. For the epilepsy example, the clinical interpretation obtained from the final Cox regression models would not have been discovered without IPD. The current author would recommend that for investigating heterogeneity, the model selection strategy should involve examination of all pre-specified clinically important treatment by covariate interactions, rather than exploring interactions only if the corresponding main effect is found to be significant. However, although unlikely in many meta-analysis situations, independent validation of interaction effects is ideally required therefore inferences should be made cautiously.

Berlin *et al*. (2002) have undertaken similar comparisons of meta-regression analyses based on IPD or AD with an empirical example of five trials. Their investigations revealed that the AD meta-regression analyses failed to detect the importance of a particular covariate, panel reactive antibodies (PRA), included as the percentage above or below a particular cut-off value. In contrast, the IPD based models revealed a clinically important and statistically significant difference between the effect of treatment among patients whose PRA value was above or below the chosen cut-off. These results show another means by which AD and IPD based meta-regression analyses could potentially differ. In addition, a simulation study undertaken by Lambert *et al*. (2002) showed that the statistical power of meta-regression using aggregate binary data was dramatically and consistently lower than that of the corresponding IPD analysis with little agreement between the parameter estimates obtained from the two methods. Both publications recommend that IPD should be used whenever feasible (Berlin *et al*. 2002; Lambert *et al*. 2002) for a reliable exploration of heterogeneity.

For the empirical comparison presented in the current paper involving a small number of trials, but still reflective of many meta-analyses in practice, the results suggest that meta-regression using AD can be accurate if there is evidence for a within study treatment by covariate interaction and sufficient between trial variation for the aggregate value of the covariate. Departures from this condition could mean that meta-regression results using AD are unreliable.

Further comparisons between IPD and AD and a systematic assessment of the empirical evidence are needed in order to provide guidelines of how, and in which situations, IPD is most beneficial for meta-analysis and meta-regression. A systematic review of empirical comparisons for the main treatment effect (Clarke *et al*. 2001) and an international collaborative effort to perform empirical comparisons of meta-regressions (J. Berlin, pers comm., ESTEEM project) are currently planned. The comparison and discussion presented in this paper can be added to this growing body of empirical evidence evaluating the benefits of using IPD or AD.

The current author agrees with the recommendations of Berlin *et al*. (2002) and Lambert *et al*. (2002) that IPD should be used whenever possible to reliably study patient characteristics and investigate heterogeneity. This recommendation is especially important when the number of trials in the meta-analysis is small and AD approaches are likely to become increasingly more uncertain. Furthermore, if time-to-event outcomes are of interest, IPD can be extremely valuable as a result of limitations reporting appropriate summary data.

Acknowledgements

We are grateful to the Epilepsy Monotherapy Trialists' Group for providing us with the data for this comparison. Catrin Tudur Smith is funded by grants from the NHS HTA.

References

Altman D.G., De Stavola B.L., Love S.B. & Stepnieweska K.A. (1995) Review of survival analyses published in cancer journals. *British Journal of Cancer* **72**, 511–518.

Berlin J.A., Santanna J., Schmid C.H., Szczech L.A. & Feldman H.I. (2002) Individual patient versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in Medicine* **21**, 371–387.

Chalmers I. (1993) The Cochrane Collaboration: preparing, maintaining and disseminating systematic reviews of the effects of health care. *Annals of the New York Academy of Science* **703**, 156–165.

Clarke M., Stewart L., Tierney J. & Williamson P. (2001) Individual patient data meta-analyses compared with meta-analyses based on aggregate data [Protocol for Cochrane review]. *The Cochrane Library*. Update Software, Oxford.

Heller A.J., Chesterman P., Elwes R.D.C., Crawford P., Chadwick D., Johnson A.L. & Reynolds E.H. (1995) Phenobarbitone, phenytoin, carbamazepine, or sodium valproate for newly diagnosed adult epilepsy: a randomised comparative monotherapy trial. *Journal of Neurology, Neurosurgery, and Psychiatry* **58**, 44–50.

Lambert P.C., Sutton A.J. & Jones D.R. (2002) A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology* **55**, 86–94.

Marson A.G., Williamson P.R., Hutton J.L., Clough H.E. & Chadwick D.W. (2000) Carbamazepine versus valproate monotherapy for epilepsy (Cochrane Review). *The Cochrane Library*, Issue 3. Update Software, Oxford.

Mattson R.H., Cramer J.A. & Collins J.F. (1992) A comparison of valproate with carbamazepine for the treatment of complex partial seizures and secondarily generalized tonic-clonic seizures in adults. *New England Journal of Medicine* **327**, 765–771.

Parmar M.K.B., Torri V. & Stewart L. (1998) Extracting summary statistics to perform meta-analysis of the published literature for survival end-points. *Statistics in Medicine* **17**, 2815–2834.

Richens A., Davidson D.L.W., Cartlidge N.E.F. & Easter D.J. (1994) A multicentre comparative trial of sodium valproate and carbamazepine in adult onset epilepsy. *Journal of Neurology, Neurosurgery, and Psychiatry* **57**, 682–687.

de Silva M., MacArdle B., McGowan M., Hughes E., Stewart J., Neville B.G.R., Johnson A.L. & Reynolds E.H. (1996) Randomised comparative monotherapy trial of phenobarbitone, phenytoin, carbamazepine, or sodium valproate for newly diagnosed childhood epilepsy. *Lancet* **347**, 709–713.

Stewart L. & Clarke M. (1995) Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine* **14**, 2057–2079.

Thompson S.G. (1994) Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* **309**, 1351–1355.

Thompson S.G. & Sharp S.J. (1999) Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* **18**, 2693–2708.

Tudur Smith C. (2004) Individual patient data meta-analysis with time-to-event outcomes. PhD Thesis. University of Liverpool, Liverpool, UK.

Tudur C., Williamson P.R., Khan S. & Best L.Y. (2001) The value of the aggregate data approach in meta-analysis with time-to-event outcomes. *Journal of the Royal Statistical Society Series A* **164**, 357–370.

Tudur Smith C., Williamson P.R. & Marson A.G. (2005) Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics in Medicine* (In press).

Verity C.M., Hosking G. & Easter D.J. (1995) A multicentre comparative trial of sodium valproate and carbamazepine in paediatric epilepsy. *Developmental Medicine and Child Neurology* **37**, 97–108.

Williamson P.R., Clough H.E., Hutton J.L., Marson A.G. & Chadwick D.W. (2002a) Statistical issues in the assessment of the evidence for an interaction between factors in epilepsy trials. *Statistics in Medicine* **21**, 2613–2622.

Williamson P.R., Marson A.G., Tudur C., Hutton J.L. & Chadwick D.W. (2000) Individual patient data meta-analysis of randomized anti-epileptic drug monotherapy trials. *Journal of Evaluation in Clinical Practice* **6**, 205–214.

Williamson P.R., Tudur Smith C., Hutton J. & Marson A.G. (2002b) Aggregate data meta-analysis with time-to-event outcomes. *Statistics in Medicine* **21**, 3337–3351.