Research Article

Statistics in Medicine

Received 8 October 2009 Accepted 26 May 2011 Published online 22 September 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4333

A new strategy for meta-analysis of continuous covariates in observational studies

Willi Sauerbrei[a*†] **and Patrick Royston**[b]

When several studies are available, a meta-analytic assessment of the effect of a risk or prognostic factor on an outcome is often required. We propose a new strategy, requiring individual participant data, to provide a summary estimate of the functional relationship between a continuous covariate and the outcome in a regression model, adjusting for confounding factors. Our procedure comprises three steps. First, we determine a confounder model. Ideally, the latter should include the same variables across studies, but this may be impossible. Next, we estimate the functional form for the continuous variable of interest in each study, adjusted for the confounder model. Finally, we combine the individual functions by weighted averaging to obtain a summary estimate of the function. Fractional polynomial methodology and pointwise weighted averaging of functions are the key components. In contrast to a pooled analysis, our approach can reflect more variability between functions from different studies and more flexibility with respect to confounders. We illustrate the procedure by using data from breast cancer patients in the Surveillance, Epidemiology, and End Results Program database, where we consider data from nine individual registries as separate studies. We estimate the functional forms for the number of positive lymph nodes and age. The former is an example where a strong prognostic effect has long been recognized, whereas the prognostic effect of the latter is weak or even controversial. We further discuss some general issues that are found in meta-analyses of observational studies. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: prognostic research; risk assessment; meta-analysis; regression modelling; fractional polynomials; confounder model

1. Introduction

When data from several studies are available, a meta-analytic assessment of the effect of a risk or prognostic factor on an outcome is often required. In observational studies, potentially with the influence of several other factors, such assessments are best done in a multivariable framework. In addition, if the factor of interest is measured on a continuous scale, a dose–response relationship must be estimated. Methodological development of each of the three elements has a long history. However, how best to combine them in a strategy suitable for the meta-analysis of dose–response functions from observational data is debatable. A key issue of the analysis is the availability of data from each study. Mostly, aggregate data from publications are the only source of information. Methods to summarize such data were proposed nearly two decades ago (e.g. [1]), and different types of applications can be found in the literature (e.g. [2], [3]). White [2] claimed to present reliable quantitative estimates of the effect of alcohol on all-cause mortality using only published data. In a letter [4], however, we raised issues of the choice of adjustment factors, different measurements of the risk factor and appropriateness of the dose–response analysis. We concluded that 'we believe that any useful quantitative summary of observational studies requires data on individuals'. In his response [5], White agreed that an ideal approach requires individual participant data (IPD).

[a] *IMBI, Freiburg University Medical Center, Stefan-Meier-Str. 26, 79100, Freiburg, Germany*
[b] *Hub for Trials Methodology Research, MRC Clinical Trials Unit and University College London, Aviation House, 125 Kingsway, London, WC2B 6NH, UK*
**Correspondence to: Willi Sauerbrei, IMBI, Freiburg University Medical Center, Stefan-Meier-Str. 26, 79100, Freiburg, Germany.*
[†] *E-mail: wfs@imbi.uni-freiburg.de*

In recent years, several IPD datasets that would support more detailed analysis have been assembled. However, strategies may ignore several aspects of the data and oversimplify the analysis. Researchers may pool the data and ignore the fact that the data were collected in separate studies. Furthermore, instead of trying to estimate a summary dose–response function, which in our view should be the aim with continuous variables, they may categorize such predictors into several groups. For example, for three groups, they may take the 25th and 75th percentiles as cut points. With the middle group as the comparator, they may give estimated effects for the low and the high groups relative to the middle group, with or without adjustment for other factors. However, more flexible statistical methodology could further improve the analysis and may provide additional insights.

Despite the cost and effort involved, collaborative projects in which several studies are assembled are becoming popular. The aim is to assess the effect of a prognostic factor, risk factor or more generally a variable of interest on an outcome. Many such variables are continuous. Some projects of this type are described in References [6–9]. Large sample size is appropriately considered essential to obtaining a reliable answer, but the analysis strategy often has weaknesses of the types mentioned in the preceding paragraphs. In a comprehensive review of recent developments in meta-analysis, Sutton and Higgins [10] discuss many topics and cite 281 references. None of these papers, however, appears to concentrate on a meta-analysis of a continuous covariate without resorting to cut points.

In this paper, we propose a new strategy for assessing the influence of a continuous variable on an outcome in a regression model by combining dose–response functions from several observational studies. For such studies, adjustment for confounding factors is almost always necessary. Our procedure comprises three steps. First, we determine a confounder model (which may differ across studies, depending on which variables are available); next, we estimate the functional form for the continuous variable of interest in each study, adjusted for the confounder model; and finally, we combine the individual functions by weighted averaging. Fractional polynomial methodology [11–13] and pointwise weighted averaging of functions are the key components. Variants of each of the three steps are possible. Under ideal conditions regarding the available confounders, the simplest approach has some resemblance to the pooled approach [9].

Our approach can be used in most situations with IPD, provided the measurements in the individual studies have a common scale. For illustration purposes, we use a large database of prognostic factors and outcomes for breast cancer patients from the Surveillance, Epidemiology, and End Results (SEER) programme [14]. The aim of the analysis was not to answer any specific clinical question. We consider data from individual registries as separate studies. We estimate the functional forms for the number of positive lymph nodes and age. The former is an example where a strong prognostic effect has long been recognized, whereas the prognostic effect of the latter is weak or even controversial [15]. The SEER database has been used several times to assess similar types of questions but by using different analysis approaches [14, 16].

Here we propose a strategy for a meta-analysis of a continuous variable under conditions that are ideal from the analysis point of view. Many difficult and critical aspects of a meta-analysis are not addressed. Related to our three steps are the standardization of measurements to a common scale, the selection of satisfactory adjustment models and estimation of the functional form. Difficulties of standardization are well described in a project on microvessel density as a prognostic factor in lung cancer [17]. To combine data across studies, Look *et al.* [18] replaced continuous measurements by within-study ranks. We consider that to be too simple an approach with disadvantages; we prefer transformations which retain a common metric [19]. Development of adjustment models depends on the variables available in the individual studies. Pooling the data implies restriction to variables available in all studies or, in a particular study, replacement of an entire variable by imputed values. These limitations do not apply to our approach.

Ideally, a meta-analysis is pre-planned, and relevant variables are collected in the individual studies [20]. Such projects have been popular in epidemiology for more than a decade [21], but we are unaware of any such project in prognostic research. In this area, our approach would be feasible because data are available from large clinical trial groups conducting a series of trials over several years. However, difficulties may arise when trying to combine data from several trial groups. Instead of using an approach similar to that of Butcher *et al.* [6], Royston *et al.* [9] preferred to construct a prognostic model from seven European Organization for Research and Treatment of Cancer studies and validate it with data from two Medical Research Council studies. Because of differences in the data from the two centres, they decided to address a different type of question, but the data can also be analysed using the methodology presented here.

We discuss several related issues to address the use and usefulness of our analysis strategy in the broader context of a meta-analysis for continuous variables. A number of additional issues in a meta-analysis are raised in the context of multiple epidemiological studies by the 'Pooling Project of Prospective Studies of Diet and Cancer' [22] and the 'Emerging Risk Factors Collaboration' group [23] (see also http://www.phpc.cam.ac.uk/MEU/ERFC/).

The plan of the paper is as follows. In Section 2, we describe the SEER database in breast cancer. Section 3 gives details of our proposed analysis strategy. Section 4 gives the results of the analyses. In Section 5, we discuss some general issues that are found in meta-analysis. Section 6 is a discussion.

2. Data

The SEER Program of the USA collected data about the incidence of cancer and related matters from 11 population-based registries (see http://seer.cancer.gov). Here we use the data extracted by Tai *et al.* [14] for modelling the effect of age in T1–2 breast cancer patients. They used data from nine registries: San Francisco-Oakland, Connecticut, Metropolitan Detroit, Hawaii, Iowa, New Mexico, Seattle (Puget Sound), Utah and Metropolitan Atlanta, which are considered as nine different studies here. See Table I for details of the registries.

Selected patients were women without previous history of cancer presenting with a non-inflammatory invasive breast carcinoma, histologically confirmed pT1–2 pN0-1 M0 and diagnosed between 1988 and 1997 in whom curative surgery and axillary lymph node dissections were performed. Certain records were rejected because of data quality concerns: uncertain sequence of treatment, non-hospital based data records, month of diagnosis unknown and race unknown. Following examination of outliers, we further excluded one case with 75 nodes involved. Here we consider death from breast cancer as the only event of interest. Death from other causes are handled as a censored survival time. Follow-up cutoff date was 31 December 1999 as provided by the database. Table II gives an overview of patient characteristics used in the analyses and their distribution. In contrast to usual meta-analyses, each of the nine 'studies' has a large sample size, and the same variables are available. To remove the problem of missing values, we created one imputation of the unknown values of estrogen receptor (ER) (31%), progesterone receptor (PR) (32%) and grade (29%). Because the analysis is mainly for illustration and to avoid unhelpful complexity, we did not use multiple imputation.

To illustrate our approach, we used the number of positive nodes as an example of a strong prognostic factor and age as a factor that is weakly prognostic [15]. Based on results from many studies on these two factors, we expect a rather stable effect of the number of nodes and an unstable effect of age. During the recruitment period, nodal status and other characteristics were often used to determine aspects of the treatment strategy. We included as adjustment factors the type of surgery and radiotherapy, with three dummy variables representing the four combinations of these treatments. The approach is computationally simpler than a stratified analysis and gives the same result under the proportional hazards assumption.

Table I. Details of the SEER registry data for primary breast cancer (T-stages 1 and 2). Follow-up is median time (years).

Study		Persons	Events	(%)	Follow-up
1	San Francisco-Oakland	14, 213	1, 270	(8.9)	6.0
2	Connecticut	11, 339	1, 084	(9.6)	6.1
3	Metropolitan Detroit	13, 533	1, 540	(11.4)	6.3
4	Hawaii	3, 666	235	(6.4)	6.0
5	Iowa	12, 028	1, 269	(10.6)	6.3
6	New Mexico	4, 422	400	(9.1)	5.8
7	Seattle (Puget Sound)	13, 671	1, 184	(8.7)	6.1
8	Utah	4, 009	386	(9.6)	6.1
9	Metropolitan Atlanta	6, 923	731	(10.6)	6.0
All		83, 804	8, 099	(9.7)	6.1

Table II. Basic description and univariate Cox regression analysis of the prognostic factors in the SEER primary breast cancer dataset. Data from all nine registries have been pooled.

Variable		% or mean (sd)	$\widehat{\beta}$	SE	$\widehat{\beta}$/SE
Race	White/other	93.1	0	–	–
	Black	6.9	0.63	0.04	17.0
Histology	Non-ductal	23.2	0	–	–
	Ductal	76.8	0.27	0.03	9.5
Marital status	Other	39.6	0	–	–
	Married	60.4	−0.15	0.02	−6.6
Tumour size (mm)		18.7 (10.1)	0.054	0.001	61.0
Breast-conserving surgery	No	60.7	0	–	–
	Yes	39.3	−0.62	0.03	−23.1
Grade	1	14.8	0	–	–
	2	42.9	1.27	0.08	15.6
	3	38.2	2.12	0.08	30.2
	4	4.1	2.22	0.10	28.3
ER	No	22.9	0	–	–
	Yes	77.1	−0.91	0.02	−40.2
PR	No	31.4	0	–	–
	Yes	68.6	−0.80	0.02	−35.8
Quadrant	Other	85.0	0	–	–
	Inner	15.0	0.06	0.03	2.1
Age (years)		59.8 (13.6)	−0.0078	0.0008	−9.3
Nodes positive		1.25 (3.26)	0.110	0.001	80.0
Nodes examined		15.5 (6.6)	0.0087	0.0017	5.3

3. Strategy for a meta-analysis of continuous variables

To determine an average function from $K > 1$ studies, we propose a strategy with three main steps. First, excluding the variable of primary interest, a confounder model M_i is determined for the ith study ($i = 1, \ldots, K$). The effect of the variables in these models is summarized in the 'confounder index', η_i, for each study. The variables may differ between studies. In the second step, the confounder index is used as an adjustment factor when determining the functional form for the variable of interest in each study. Based on fractional polynomials, we propose three approaches to estimating the functional form. In the third step, a weighted average function is calculated. The strategy is illustrated in Box 1.

3.1. Determination of confounder models

In general, a summary estimate of the function based on univariate analyses can only be considered as a starting point. A comprehensive assessment requires adjustment for other potential confounding factors in a model. Sauerbrei and Royston [12] and Royston and Sauerbrei [13] proposed the multivariable fractional polynomial (MFP) approach.

Denote Z as the main variable of interest and $X_1, \ldots, X_{k_i}$ as potential prognostic factors in study i. For a discussion of selecting potential factors, see Section 5.2. In general, different prognostic variables are likely to be available in individual studies, but some established predictors (e.g. in cancer, age, sex, tumour size, number of positive nodes, metastasis) should be available in each study. The first step is to determine the confounder model M_i in study i. Ignoring Z, we determine M_i by applying MFP to the variables in the ith study, the result of which is summarized in the confounder index η_i. According to preference, MFP may be replaced by another model-building strategy.

Alternatively, when the same confounders are available (or are chosen) in each study, the model may be selected by applying MFP to the pooled dataset, stratifying by study. This approach may be preferred when some of the studies are small; otherwise, power will be low and the results correspondingly unstable. If a well-established prognostic index is available, such as the Nottingham Prognostic Index in breast cancer, it may be used as the confounder index η_i without further selection or modification.

Step 1. Derive confounder models

Determine confounder models per study, $M_1, \ldots, M_k$. Summarize in corresponding indexes (linear predictors) $\eta_1, \ldots, \eta_K$.

Step 2. Determine functional form.

We propose three methods for determining the functional form of a continuous predictor in each study. In all the approaches, estimates are adjusted for the confounder models found in step 1.

1. Overall fractional polynomial (FP). Using the function selection procedure (FSP), find the best FP transformation for the pooled data set (stratified by data set). Select using a small nominal significance level.
2. Studywise FP2. Select the best FP2 function for each study.
3. Studywise selected FP. Use the FSP to select the best FP for each study individually.

Step 3. Averaging the functions

We propose three methods for determining an average function. In each method, estimates are adjusted using the confounder indexes $\eta_1, \ldots, \eta_k$ determined in step 1.

1. Pooled function
2. Function estimated with fixed-effects weights
3. Function estimated with random-effects weights

Box 1. Three steps of the meta-analysis procedure to determine an adjusted averaged dose–response function from several studies

3.2. *Determination of functional form for each study*

Before exploring functional relationships for Z, it is essential that measurements of Z be comparable across studies. Particularly for continuous variables, it may be necessary to standardize measurements as a preliminary step. See Section 5.3 for further comments. In the following, we assume that Z is measured on the same scale in all studies.

Adjusting for the index η_i of model M_i, the influence of Z in study i is estimated by the best FP2 function. Alternatively, the FSP could be used—see Appendix Section. However, the FSP may often result in a linear function because the power to detect non-linearity is too small. Therefore, small power will introduce additional variability of the dose–response functions.

Whether the functions can be averaged as described below should be decided after investigating the heterogeneity of the functional forms. A plot of all functions gives a graphical assessment. The plot may indicate that averaging is not meaningful because the functional forms differ too much between studies. We admit that the approach is heuristic, without firm methodological underpinning.

3.2.1. Method 1: Overall FP. Adjusting for the indexes $\eta_1, \ldots, \eta_K$ determined in step 1, we select the FP function that best fits the joint data according to the FSP for a given α level. As the joint data set will be large in most cases, it will have substantial power to find even weaker non-linear effects of Z. Therefore, we propose to use a small nominal significance level (e.g. 0.01 or 0.001) to determine the FP function. The same selected FP powers are used in each individual study. The adjusted FP2 model in the ith study is

$$\beta_{1i} Z^{p_1} + \beta_{2i} Z^{p_2} + \gamma_i \eta_i, \text{ or } \beta_{1i} Z^{p_1} + \beta_{2i} Z^{p_1} \ln Z + \gamma_i \eta_i \text{ when } p_1 = p_2$$

In what follows, the 'repeated powers' case $p_1 = p_2$ is implicit in the standard FP2 formula. Fixing the powers in this way, we avoid the most unstable situation of flexible modelling leading to serious artefacts in fitted functions in small studies. Individual curves are determined by estimating β_{1i} and β_{2i}.

3.2.2. Method 2: Studywise FP2. Adjusting for $\eta_1, \ldots, \eta_K$, we select the best-fitting FP2 function in each study. The adjusted FP2 model in the ith study is

$$\beta_{1i} Z^{p_{1i}} + \beta_{2i} Z^{p_{2i}} + \gamma_i \eta_i$$

Compared with the overall FP method, we allow more flexibility in the selected functions. The disadvantage is to increase the instability of the functions.

Variants of the strategy are possible. For example, if the function is required to be monotonic, restriction to FP1 functions may be more suitable.

3.2.3. Method 3: Studywise selected FP. As before we adjust for $\eta_1, \ldots, \eta_K$, but in contrast to the Studywise FP2 method, we determine the FP separately in each study, allowing the FSP to select a simpler FP1 or linear function in each study. Depending on the statistical power and the strength of the relationship with Z, the selected function in study i may be linear, FP1 or FP2:

$$\text{Linear: } \beta_{1i} Z^1 + \gamma_i \eta_i$$

or

$$\text{FP1: } \beta_{1i} Z^{p_{1i}} + \gamma_i \eta_i$$

or

$$\text{FP2: } \beta_{1i} Z^{p_{1i}} + \beta_{2i} Z^{p_{2i}} + \gamma_i \eta_i$$

Smaller studies will have only low power to find non-linear effects. Therefore, we propose to use larger nominal significance levels (e.g. 0.05 or 0.157, the latter being similar to an Akaike Information Criterion (AIC)). Smaller significance levels may often result in choosing a linear function in several studies. Because of limited statistical power, determining studywise FP functions will increase the instability of selected functions.

3.3. Averaging of functions

Since the Cox model has no intercept term (risks or hazards are estimated relative to an unspecified baseline hazard function), each estimated function must first be standardized. If $\widehat{f}_i(Z)$ is the estimated function in the ith study, the standardized function is

$$\widetilde{f}_i(Z) = \widehat{f}_i(Z) - \widehat{f}_i(Z_0)$$

where Z_0 is a specific value of Z chosen as a suitable reference point for all studies. For example, if Z is age, Z_0 may be 50 or 60 years. For nodes, $Z_0 = 0$ (i.e. node-negative patients) is a natural choice. The function $\widetilde{f}_i(Z)$ is an estimate of the effect of Z in study i in relation to the reference point.

As an estimate of the overall function, we average $\widetilde{f}_i(Z)$ with an appropriate choice of weights:

$$\widetilde{f}(Z) = \sum_{i=1}^{K} w_i(Z)\, \widetilde{f}_i(Z),$$

where $w_i(Z)$ are weights based on the variance of the $\widetilde{f}_i(Z)$ in individual studies. Fixed or random effects assumptions lead to different variance structures and correspondingly different weights.

3.4. Estimation for fixed and random-effects models

Suppose we have K studies with effect estimates ψ_i and $\text{var}(\psi_i) = v_i$, $i = 1, \ldots, K$. In the present context, ψ_i and v_i, and all the quantities derived from them, are functions of the covariate Z of interest and are calculated pointwise. For simplicity, the argument (Z) is suppressed in what follows, but note that all estimates and variances, specifically Ψ, Ψ^{rand}, τ^2, Q, ψ_i, v_i, w_i and w_i^*, are indexed by Z. The formulae used below are taken from Sterne and Sharp's Stata program META (StataCorp, College Station, Texas, USA) [24].

3.4.1. Fixed effects. All sums are between studies, i.e. over $i = 1, \ldots, K$. Let $W = \sum v_i^{-1}$ be the sum of the fixed-effects weights. Define standardized weights as $w_i = v_i^{-1}/W$. The overall estimate and its variance are

$$\Psi = \sum w_i \psi_i, \text{var}(\Psi) = W^{-1}$$

3.4.2. Random effects. Let $W^* = \sum \left(v_i + \tau^2\right)^{-1}$ be the sum of the random-effects weights, where τ^2 is the component of variance of ψ between studies. Define standardized weights as $w_i^* = \left(v_i + \tau^2\right)^{-1} / W^*$.

The overall estimate and variance are

$$\Psi^{\text{rand}} = \sum w_i^* \psi_i, \text{var}\left(\Psi^{\text{rand}}\right) = \left(W^*\right)^{-1}$$

To estimate τ^2, let $Q = \sum v_i^{-1} \left(\psi_i - \Psi\right)^2$. Then

$$\tau^2 = \max\left\{0, \frac{Q - (K-1)}{\sum v_i^{-1} - \left(\sum v_i^{-2}\right) / \left(\sum v_i^{-1}\right)}\right\}$$

4. Results

We use the strategies described above to derive functions for the effect of the number of positive nodes and age on the relative hazard for overall survival. In two separate analyses, nodes and age assume the role of Z in turn, with different confounder models. Separately for nodes and age, we first determine the confounder models $M_1, \ldots, M_K$ and their indexes $\eta_1, \ldots, \eta_K$. We determine the functional forms according to the three different methods. We then perform a fixed-effects meta-analysis of the effects of nodes and age and illustrate the differences resulting from a random-effects approach. Finally, we carry out some sensitivity analyses.

4.1. Confounder models for nodes and age

For the pooled data, we give in Table II the distribution and the size of the prognostic effect in a univariate analysis for all available variables. Linearity was assumed for continuous variables. Because of the large sample size, all effects are highly significant, the only exception being quadrant which just reaches significance at the 5% level.

Table III gives the variables selected by MFP for the adjustment models for nodes and age. Age, nodes, nodes examined and tumour size are continuous variables. Except for nodes and nodes examined, all variables were candidates for the adjustment model for nodes, and all variables except for age were candidates for the adjustment model for age. The three dummy variables representing the combination of type of surgery and radiotherapy were forced into each model. For both adjustment models, the tumour size, grade and the ER/PR combination were selected in each study. Following preliminary investigation of its functional form, tumour size (`size`) was transformed using the sigmoid function

$$f\,(\texttt{size}) = \Phi\left[0.074(\texttt{size} - 18.67) + 0.601\right]$$

where $\Phi\,(.)$ is the standard normal distribution function. The function $f\,(\texttt{size})$ flattens off at large values of `size` and is consistent with the functional form found by Vinh-Hung *et al.* [16]. As expected, FP functions selected for $f\,(\texttt{size})$ were mostly linear.

Only in five studies is age included in the confounder model for nodes, and the selected functions differ substantially. This indicates that age is not a strong prognostic factor. In contrast, nodes is included in each of the adjustment models for age, and in five of the nine studies a log transformation is selected.

4.2. Weights for averaging functions

We illustrate the dependence of variances and their respective weights on Z and the differences between the weights appropriate to fixed and random effects models. We first explain how the various quantities in the calculation relate to one another.

Consider for example age 40 years (i.e. $Z = 40$) for three of the nine studies (1, 4 and 5). The study-wise selected FP method was used. Table IV shows for $i = 1, 4, 5$ the value ψ_i of the fitted function, its estimated within-study variance v_i, the fixed-effects weight $w_i = v_i^{-1} \big/ \sum v_i^{-1}$ and the random-effects weight $w_i^* = \left(v_i + \tau^2\right)^{-1} \Big/ \sum \left(v_i + \tau^2\right)^{-1}$. For example, at age 40 years, the fixed-effects weight for study 1 is calculated as

$$w_1 = \frac{0.0031^{-1}}{0.0031^{-1} + 0.0100^{-1} + 0.0037^{-1}} = 0.47$$

Table III. Confounder models selected by MFP for each study. (a) Nodes, (b) age. 'x' means the (binary) variable was selected. Numbers denote FP powers, 1 meaning a linear function was selected.

Variable	Study								
	1*	2*	3*	4	5*	6	7*	8	9
(a) Nodes									
Race	x		x				x		x
Histology	x	x	x	x	x		x		x
Marital status	x	x	x						
Size**	0	1	1	−2, −2	1	1	1	1	1
BC surgery***	x	x	x	x	x	x	x	x	x
Grade	x	x	x	x	x	x	x	x	x
ER/PR	x	x	x	x	x	x	x	x	x
Quadrant		x					x		
Age	0.5, 2	1	−2, −2		2, 3		1		
(b) Age									
Race	x	x	x				x		x
Histology	x	x	x		x		x		x
Marital status	x	x	x						
Size**	0	1	1	1	1	1	1	1	1
BC surgery***	x	x	x	x	x	x	x	x	x
Grade	x	x	x	x	x	x	x	x	x
ER/PR	x	x	x	x	x	x	x	x	x
Quadrant	x	x	x		x	x	x		x
Nodes	−0.5, 3	0	−2, 0.5	−0.5	0	0	0	0.5	0
Nodes ex.	1	1	1		0	1	−1, 3	1	1

*Studies with more than 1000 events.
**Variable transformed.
***Treatment variable, forced to be included in the model.

Table IV. Illustration of quantities required in the calculation of an overall effect estimate in a meta-analysis. See text for details.

Study	Effect estimate	Variance	Weight	
			Fixed	Random
i	$\widehat{\psi}_i$	v_i	w_i	w_i^*
1	0.51	0.0031	0.47	0.36
4	0.11	0.0100	0.14	0.29
5	0.41	0.0037	0.39	0.35

The overall estimated function at age 40 using fixed effects weights is

$$\widehat{\Psi}(40) = 0.51 \times 0.47 + 0.11 \times 0.14 + 0.41 \times 0.39 = 0.41$$

The w_i and the w_i^* each sum to 1 over the three studies. They assess the relative contribution of each study's estimated function to the weighted average function across studies. We see that with fixed-effects weights, studies 1 and 5 dominate in the overall estimate at age 40 years, whereas with random-effects weights, the relative contribution of the three studies is similar.

For the same three studies (1, 4 and 5), Figure 1(a) shows the FP functions for age selected using studywise selected FP. We chose 60 years as the reference value. The functions show the dependence on age on a log hazard scale. FP2 functions are selected in the two large studies (1 and 5), whereas a linear function is selected in the small study. As in the two large studies, a U-shaped function also fits better in study 4, but because of low power, a linear function is selected.

The within-study variance is shown for the three functions in Figure 1(b). The variance is (forced to be) zero at the reference point 60 years and increases with distance from the reference point. The small study has a larger variance with the exception of age > 90 years. The two large studies indicate a strongly increased risk for very old women. Data are sparse in this region and the variance is large. A

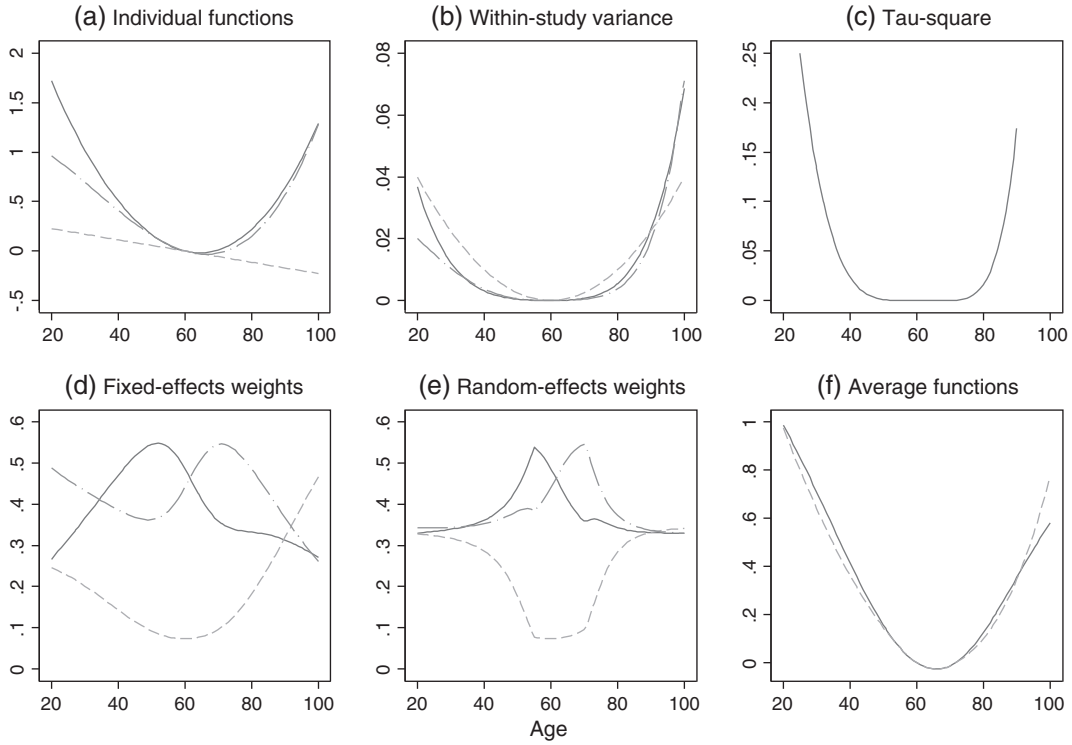

Figure 1. Illustrative example to show variance functions within (b) and between (c) studies, with resulting weights (d, e) and average functions (f). Note the different vertical scales in (b) and (c).

smaller number of patients is available in study 4, but the variance of the linear function is smaller. The fixed-effects weights for averaging the functions across studies are derived from the reciprocal of the variances (Figure 1(d)). The local maxima for studies 1 and 5 are a result of the distribution of events by age. Whereas study 1 has many more patients and events (302 vs 191) in the age group 41–50 years, study 5 has more events (272 vs 185) in the age range from 71 to 80 years. An enlarged scale would show this difference in the variances.

Figure 1(c) shows the between-study component of variance, τ^2. Individual FP functions (see Figure 1(a)) are similar in the age range from about 45 to 75 years; therefore, τ^2 is small in this range. For values outside, the linear function differs markedly from the two FP2 functions, and τ^2 rises steeply. It dominates the total variance, where the effect of differences between within-study variances becomes negligible. Therefore, the random effects models assign nearly the same weights to all studies in this age range (Figure 1(e)). Because the relative weight functions for fixed and random effect models differ substantially only in this range, the average functions are similar (Figure 1(f)).

As often happens in regression models, the choice of reference points affects the way the results appear. In our case, choosing a different reference point for Z alters the variance of the estimated curve and therefore the weight associated with each point. This in turn affects the averaged function. Therefore, the reference point should be chosen sensibly.

Restriction to three studies was done for the purpose of illustration. In what follows, we always use all nine studies.

4.3. Average function for nodes

In Figures 2 and 3, we show the functions of nodes selected with the three methods, standardized to 0 nodes. FP powers selected are given in the upper part of Table V. All functions are adjusted for the indexes derived from the confounder models shown in Table III. Figure 2 shows the functions for the individual studies, whereas Figure 3 gives the average function with 95% pointwise confidence intervals. The average functions are similar for all three methods, with a steep rise at the beginning and a flattening for more than 10 positive nodes. Differences are visible for a very large number of nodes. The functions

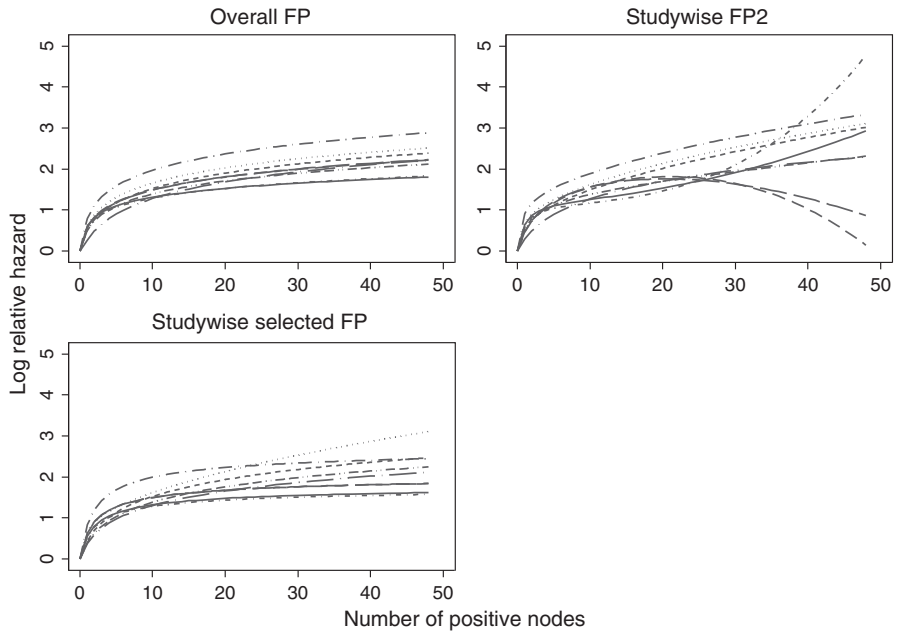

Figure 2. Individual function estimates for nodes in each study, by three estimation methods.

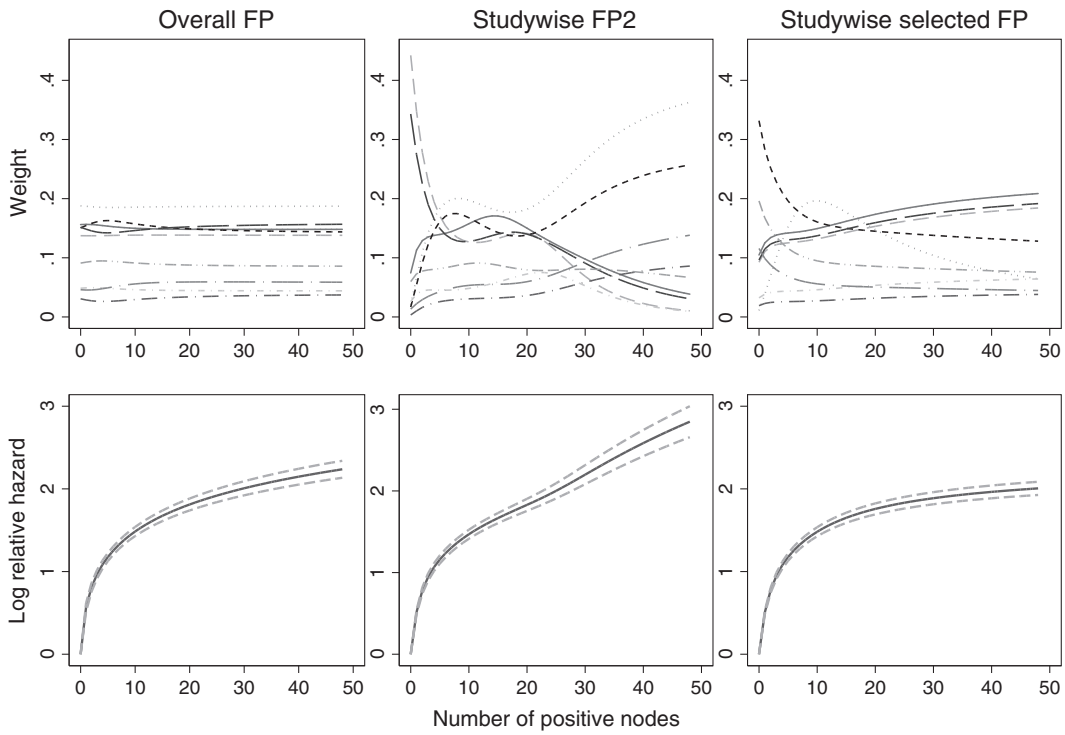

Figure 3. Upper three panels: fixed-effects weights for `nodes` in each study. Lower three panels: average function estimates, with pointwise 95% confidence intervals, by three estimation methods. Curves for function estimates are standardized to 0 at 0 nodes.

derived with method 2 increase substantially even for a larger number of nodes. However, in this region the sample size is small, which introduces considerable uncertainty. For method 2, functions in the individual studies differ for a larger number of nodes, whereas for the other methods, all individual functions have a similar shape.

Table V. FP powers obtained according to three methods of model selection (1, overall FP; 2, studywise FP2; 3, studywise selected FP). Upper part: models for nodes; lower part: models for age. All models are adjusted for the respective confounder index derived from the models shown in Table III.

Variable	Method	Study								
		1	2	3	4	5	6	7	8	9
Nodes	1	−2, 0	−2, 0	−2, 0	−2, 0	−2, 0	−2, 0	−2, 0	−2, 0	−2, 0
	2	−1, 2	0, 3	−2, 0.5	−2, 0.5	−2, 0.5	−1, 3	0, 2	−0.5, 0	−0.5, 1
	3	−0.5	−0.5	−2, 0.5	−0.5	0	−0.5	−0.5	0	0
Age	1	0.5, 1	0.5, 1	0.5, 1	0.5, 1	0.5, 1	0.5, 1	0.5, 1	0.5, 1	0.5, 1
	2	0.5, 3	1, 1	−2, −1	−2, 2	2, 3	3, 3	3, 3	3, 3	0, 0
	3	0.5, 3	1	1	1	2, 3	1	3, 3	1	1

4.4. *Average function for age*

In Figures 4 and 5, we show the functions of age selected with the three methods, standardized to age 60 years. FP powers selected are given in the lower part of Table V. Except for method 3, the average functions are broadly similar, with some differences for very young and very old women. The risk is low for women between about 40 and 65 years, and outside this range it increases. Method 3 indicates a monotonic increasing risk, a result which does not agree with current thinking about the age function (see also the discussion). For all methods, the function varies between the individual studies (Figure 4), indicating that age is not a strong prognostic factor when adjusted for other variables.

4.5. *Fixed versus random effects models*

The upper panels of Figure 6 show the component of variance, τ^2, from a random effects model, as a function of Z. For nodes, τ^2 is similar up to about 20 nodes; beyond that, methods 2 and 3 have much larger values. Only for very young and very old patients is the random effects variance substantial. In contrast to nodes, there is little difference among the three methods in the middle of the range, from about 25 to 85 years.

Up to about 25 nodes, the average functions are similar; the confidence intervals are wider for the random-effects model. For a larger number of nodes, the fixed-effect function indicates a slightly

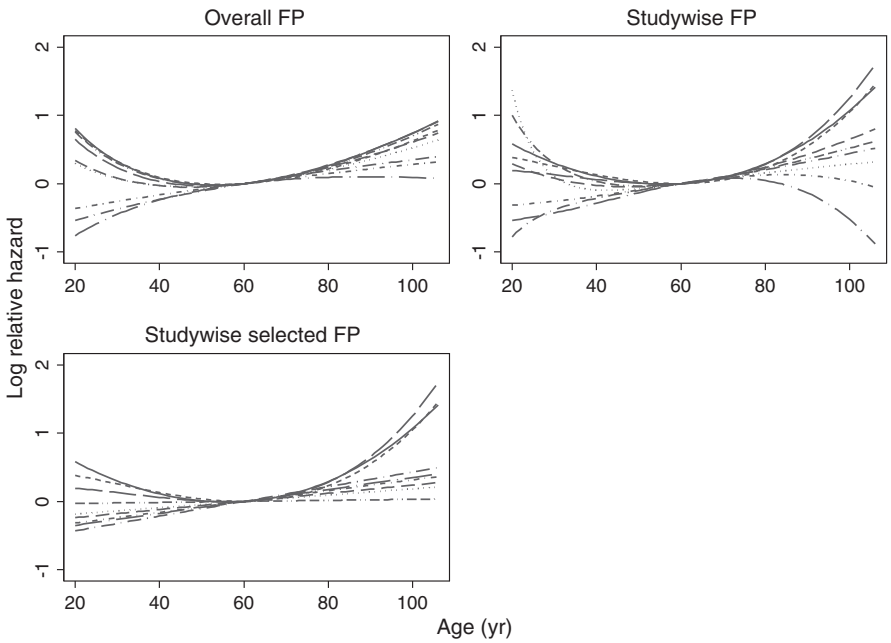

Figure 4. Individual function estimates for age in each study, by three estimation methods.

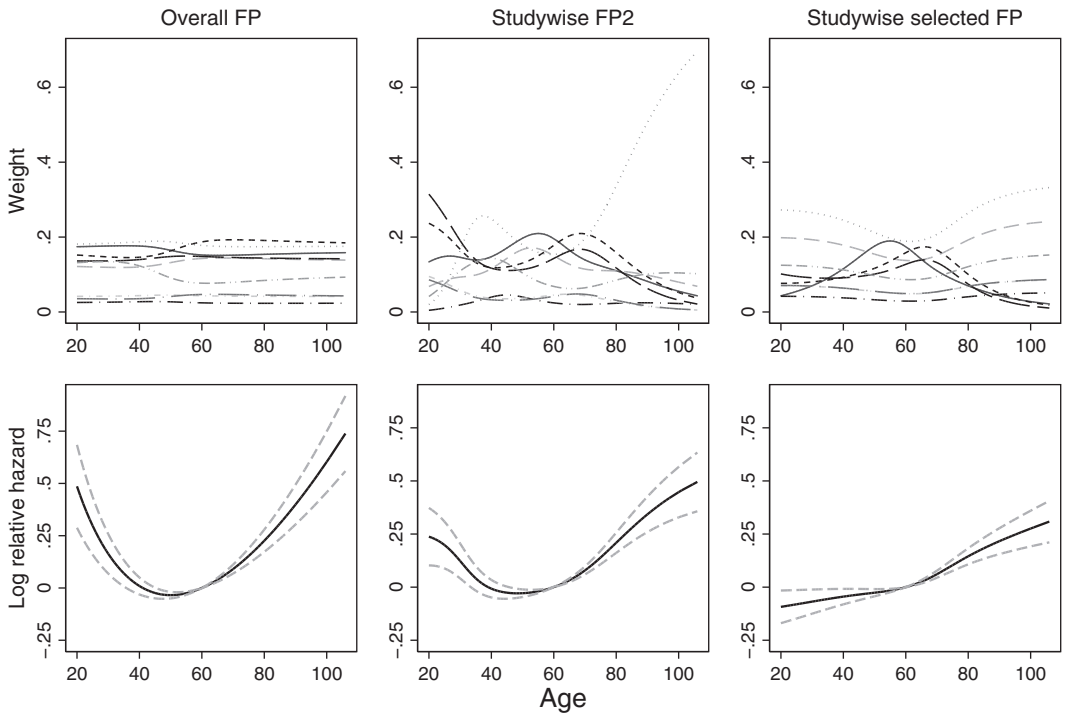

Figure 5. Upper three panels: fixed-effects weights for age in each study. Lower three panels: average function estimates, with pointwise 95% confidence intervals, by three estimation methods. Curves for function estimates using fixed-effects weights are adjusted for the confounder indexes and are standardized to 0 at age 60 years.

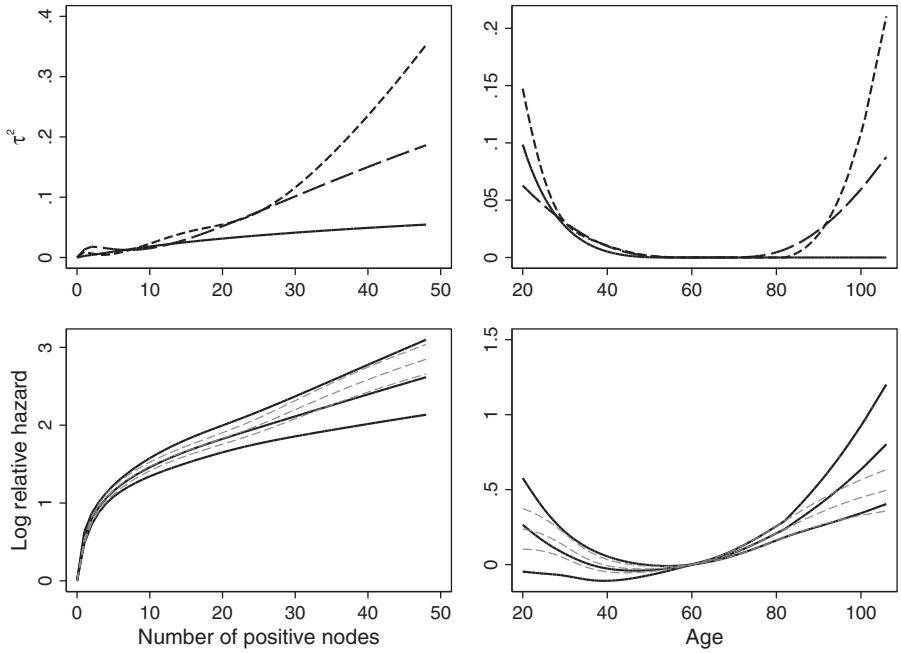

Figure 6. Random effects models. Upper panels: estimates of $\tau^2(Z)$, the random-effects variance component between studies for nodes (left) and age (right). Solid lines, overall FP; short dashes, studywise FP2; long dashes, studywise selected FP. Lower panels: average function with 95% pointwise confidence interval for the studywise FP method. Dark solid lines, random effects model; pale dashed lines, fixed-effects model.

increased risk. For patients between about 40 and 85 years, the functions and confidence intervals for age are similar, but outside this range, the random-effects intervals are much wider. The estimates from the random-effects model show a higher risk for patients over 85 years.

4.6. *Sensitivity analysis*

4.6.1. Unadjusted function. For methods 1 and 3, we present the individual figures and the averaged functions (fixed effects) from an unadjusted analysis (see Figure 7). The average Overall FP function has a shape similar to the corresponding adjusted function (Figure 5, lower panel), but the effect of age is much larger for values away from the reference value (60 years). In contrast, the functional forms of the averaged studywise selected FP functions are substantially different in the adjusted and unadjusted analyses. The former shows an increase with age, whereas the latter shows a decrease. Analogous analyses for nodes give similar results for averaged adjusted and averaged unadjusted functions (data not shown).

4.6.2. Smaller sample size. In contrast to most meta-analyses of observational studies, the sample size of the individual studies within the SEER database is large. Reflecting a more realistic situation and also to assess the effect of sample size on our procedures, Figure 8 presents results from a 10% random subsample of SEER. Results for nodes largely agree with the main analysis, whereas the average age function indicates no substantial effect. With the smaller sample size, the probability of detecting non-linearity is smaller, and a linear function is often selected, in contrast to the more realistic U-type shape.

5. General issues in meta-analysis

We have avoided several critical issues of a 'real' meta-analysis by using the data from SEER registries as individual studies. Some aspects of practical relevance are discussed in this section.

5.1. *Selection of studies and publication bias*

In our example, we used the 'studies' extracted for another project by Tai *et al.* [14]. We thereby avoided the difficult tasks of selecting relevant studies and collecting individual participant data (IPD) from each study. Most meta-analyses are based on published data, but our approach requires one to build prognostic models in each study and therefore needs IPD. Even randomized trials present serious difficulties and costs [25]. Such projects are rarely done with prognostic factors. Altman *et al.* [26] discuss several steps needed in an IPD meta-analysis of microvessel density in lung cancer. Furthermore, our approach

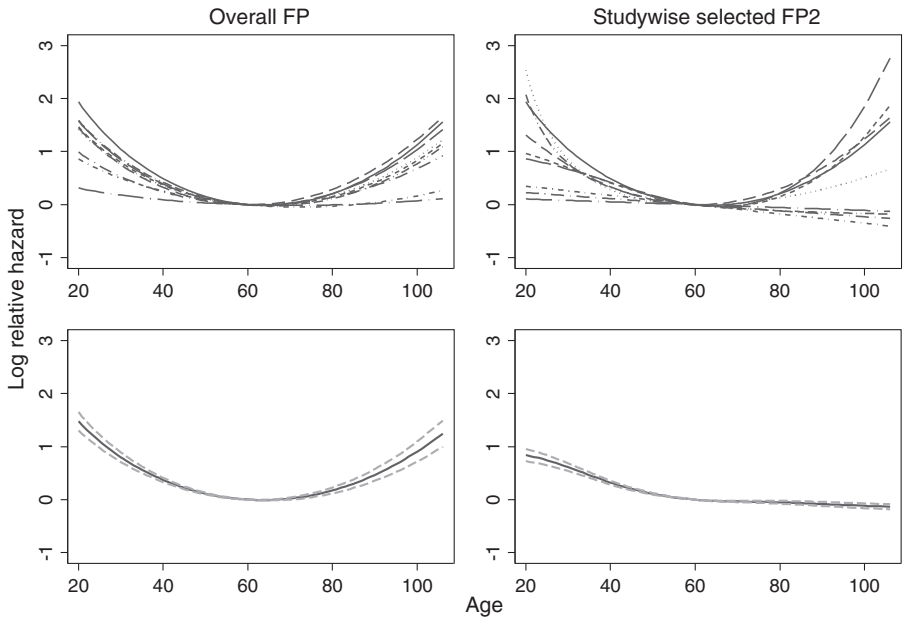

Figure 7. Functions of age, unadjusted for confounder indexes (fixed-effects models). Left panels, overall FP method; right panels, studywise selected FP method; upper panels, individual functions; lower panels, means with 95% pointwise confidence intervals, fixed-effects weights.

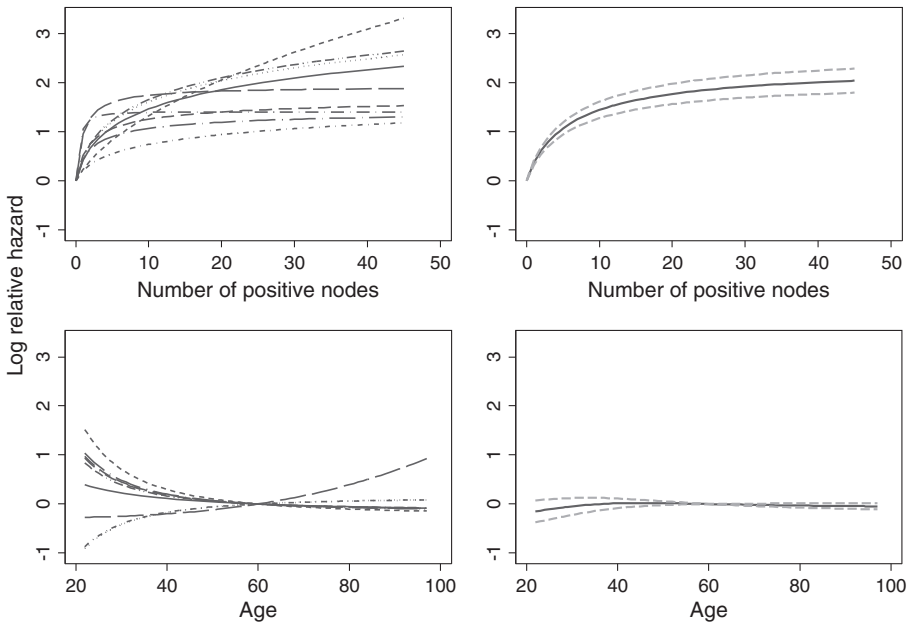

Figure 8. Ten percent random subsample of the SEER dataset. Left panels: individual functions for nodes and age, studywise selected FP method, fixed-effects model; right panels: average functions and 95% confidence intervals for nodes and age, same method.

assumes that the selected studies and their postulated functions are free of publication bias. If eligible studies were not published or did not provide IPD, the functional form for the continuous factor of interest may be affected. For a more detailed discussion and additional issues, see Reference [27].

To reduce bias, one option is to select only studies which include key confounders. For example, in a meta-analysis of published data assessing the effect of alcohol consumption on esophageal carcinoma, Rota and colleagues [28] included 'only studies that reported smoking-adjusted estimates'.

5.2. Selection of variables for the confounder models

After imputing missing data for some variables, we had complete patient data for several established prognostic factors in breast cancer. This allowed us to determine a confounder model in each study based on the available variables. The main issue was to decide whether we wanted a 'full' confounder model including all variables except Z or whether we preferred to select the variables and functions for continuous measurements. We used MFP with a significance level considered suitable to derive the confounder index. In every meta-analysis, the complexity of the chosen confounder models of course depends on the data available in the individual studies. In 'real' meta-analyses, the number of variables jointly available in all studies is often small (regarding the need for 'standardized' measurements, see the next subsection), whereas many additional variables are recorded only in subsets of the studies. To avoid residual confounding in the individual studies, it may be preferable to adjust for 'many' factors or at least for factors with a stronger effect in particular studies. However, such an approach must often result in considerable heterogeneity of the confounder models, and averaging the adjusted functions of Z from several studies may be doubtful. At the cost of increased residual confounding in individual studies, it may be preferable to adjust only for a small number of factors available in all studies. The correlation of the variable of interest with omitted variables and the strength of their effects is a consideration. The issue is discussed in an epidemiological context by Blettner *et al.* [21]. Residual confounding is also an important criterion for choosing the significance level to use when selecting the confounder model [29].

5.3. Standardization of measurements across studies

Every meta-analysis of a continuous variable Z requires Z to be measured on a common scale across all studies. This is not a critical issue for age and number of positive nodes, the two variables considered here. However, it is unlikely that a common scale is used for laboratory values in a larger number

of studies to be included in a meta-analysis. For example, in a pooled analysis to assess the effect of urokinase-type plasminogen activator (uPA) and plasminogen activator inhibitor-1 (PAI-1) in breast cancer, different measurement techniques were used, and the resulting values showed major heterogeneity between studies. Look *et al.* [18] ranked the measurements in each study and divided the ranks by the number of patients, thus converting their variables of primary interest (uPA, PAI-1) to fractional ranks between 0 and 1. Rank transformations clearly discard information, and more importantly, they lack a meaningful metric.

Royston *et al.* [19] propose median scaling and a transformation originally designed to improve robustness of FP functions as two appealing procedures for standardizing continuous variables for a meta-analysis of dose–response functions. These standardizing transformations may also be used to convert continuous confounders to a common scale. In every meta-analysis, related problems will arise for categorical or categorized continuous confounders. For example, in our meta-analysis we have just one binary variable representing negative or positive hormonal receptor status. Most likely, some patients had continuous measurements of estrogen and progesterone receptors, whereas in others only one of the receptors was measured. Furthermore, different cut points were probably used to categorize a patient as being receptor positive or negative. Often only a 'crude' such variable can be used in a meta-analysis. This will cause problems related to measurement error [23], [22].

5.4. Estimate from pooling over all studies as an estimated alternative average curve

A systematic search for relevant studies will often identify small studies to include in a meta-analysis. Deriving dose–response functions in individual studies will lack power to identify even stronger non-linearity in these functions. Averaging of individual functions is likely to produce severe bias. In such a situation, determination of only one dose–response function from all data pooled (after a standardizing transformation) by using our FSP may be preferable. Often, a combination of pooling several smaller studies and analysing larger studies separately is a way forward to handle smaller studies in a meta-analysis. Lacking common established prognostic factors, a univariate assessment of Z may be advisable, especially if many of the studies are small. However, results from univariate analyses are less useful in risk assessment and prognostic research.

6. Discussion

We propose a new approach to a meta-analysis of continuous covariates in observational studies, which comprises three main steps. We have illustrated it by assessing the prognostic effect of two continuous variables on the survival time of breast cancer patients. However, our approach is generic and can be used with simple modifications with different types of outcome variable. We summarize the steps as follows. First, a confounder index is determined in each study separately. Second, the functional form of the variable of interest is selected by using fractional polynomials, adjusting for the confounder index. Third, a weighted average function is calculated. Several variants of each step are possible, the choice being influenced by the (effective) sample size of the individual studies, the 'common' set of 'established' confounders, and other considerations. IPD are required. We illustrate the strategy with data from the SEER program by deriving the functional form for the number of positive nodes and for age as prognostic factors in breast cancer. Data from nine registries are treated as nine separate studies in a meta-analysis.

Although we are working with IPD, the example is artificial for several reasons. First, all the individual studies are large. Second, a large set of common variables is present, and no study has additional variables which would be candidates for the adjustment model. Third, all variables are measured on a common scale, so standardization of continuous variables within studies is not necessary. Fourth, publication bias is not a problem as the studies we included can be seen as a well-defined group. The importance of these issues and ways to handle them in a real meta-analysis are discussed in the preceding section on general issues in meta-analysis.

6.1. Potential applications

Although we have presented our strategy using an unrealistic example, the strategy has many potential applications in clinical epidemiology. Currently, we are using it to rework the meta-analysis of the prognostic effect of uPA and PAI-1 in breast cancer [18]. In epidemiology, risk assessment for a continuous

covariate is often done by using cut points, resulting in dose–response relationships that resemble step functions. Although the use of step functions produces unrealistic functions, discards information and can be further criticized [30, 31], it is still popular for the analysis of single studies and is also used in meta-analysis [6, 32]. If IPD are available, our approach can produce substantially better dose–response functions, which use the full information in the data and which are biologically more realistic.

To overcome the difficulty of obtaining IPD, researchers may decide to include in their protocols only 'higher quality' studies. Sample size (or number of events) would be a key quality criterion, thereby limiting the meta-analysis to larger studies. Obtaining IPD would be much easier, and the analysis could be done according to our new approach. Problems caused by small sample sizes would be much reduced, and it is unlikely that serious bias would be introduced by excluding smaller studies. Attempting to include 'all' studies usually results in a meta-analysis based on literature data. For observational studies, there are many issues that hinder obtaining reliable results [4, 20, 21, 33]. Our approach can also be used in prospectively planned meta-analyses, a concept that has been used in epidemiology for quite a long period [21, 34]. Potential confounders could be defined prospectively in the protocol, and measurement techniques could be standardized. Critical issues discussed in Section 5 could be avoided, and the analysis could be very similar to our analysis of the SEER registries.

6.2. Analysis options

We have proposed several options for all three steps of our procedure. Depending on what data are available, the confounder index may be derived with the given data and include a smaller or larger number of confounders (ideally, the same in all studies), or it may be taken from the literature. For example, the Nottingham Prognostic Index (NPI [35]) is well established in breast cancer. As the three components of the index (number of positive nodes, tumour size, grade) are available, we could have used NPI as the confounder index for deriving the age function. Given the NPI, the average age function (similar to the function in Figure 5) may have shown that age has additional prognostic value [36]. Of course, using NPI as a confounder index for the number of positive nodes would create other difficulties since categorized nodes are part of the NPI.

To estimate the dose–response relationship, we considered three functions from the FP class. Using the FSP to determine the function in each study (studywise selected FP approach) seems not to work well for smaller sample sizes and/or weaker effects. Whereas the results for nodes (with a strong effect) are similar with all three approaches, the function for age (weak effect) differs and is not consistent with medical knowledge. Despite a large sample size in each study, a linear or FP1 function is selected in some studies, whereas a non-monotonic FP2 function seems more plausible. An FP2 function is estimated with both the overall FP and the studywise FP methods. Whereas the former determines the FP2 power terms only once in the pooled data, the latter determines them separately in each study. The former is less flexible and may ignore much of the uncertainty in the functional form; the latter is more flexible and fits the data better but at the cost of increased instability. For more than about 20 nodes, studywise FP selects unrealistic functions in some studies, whereas for the bulk of the data, the individual functions and the average function are similar. For more extreme values in the left and right tails, overall FP also suppresses some variation in the age functions. We speculate that the average function for age from overall FP overestimates the effect of age on very young and very old women. For our outcome measure of overall survival, we consider the function from studywise FP to be more realistic.

To incorporate background knowledge, e.g. a U-shaped function for age, the class of permissible FP functions could in principle be restricted. Alternatively, instead of FP functions, it would be possible to use spline functions. They are more flexible but more unstable. Since guidance on which approach to use is absent, we prefer modelling in the FP framework. To handle issues of unrealistic functions for extreme values (e.g. some of the individual functions for more than 30 nodes derived with studywise FP), a robustness transformation may be useful [37].

After deciding on a suitable reference point, function averaging is straightforward. The final decision is whether to use weights appropriate to a random-effects or fixed-effects model. In meta-analyses of a point estimate, the random effects approach gives wider confidence intervals, but point estimates are usually similar. Beside the wider interval, we also see an effect on the functional relationship at both ends of the range of Z. For example, the function from the random-effects model for age is less extreme than that from the fixed-effects model. As all functions are derived in a data-dependent fashion, the resulting confidence intervals are too narrow. More realistic estimates may be obtained by incorporating model uncertainty [38], [39].

6.3. Bias

Bias arising from different sources is always an issue of concern in observational studies. In our context, the use of different confounder models in different studies may bias the averaged function. Different populations and unidentified subgroup effects within studies are among other sources of potential bias. It is fair to say, therefore, that we are summarizing possibly biased estimates of functions across several studies.

In observational studies, model building is usually required and will introduce its own bias. The question is whether the bias is severe and how much it matters. Sensitivity analyses may give an impression of its severity and may sometimes lead us to modify the summary function. To avoid model-building bias, we could pre-specify our analysis, but we do not think the necessity of imposing such heavy restrictions would make it an acceptable alternative to a model-based approach. It could even make the bias worse. Identifying and modelling sources of potential bias is gaining importance in the analysis of single studies and in meta-analysis [40], [41]. In contrast to our rather artificial example using the SEER dataset, many generic issues arise in the meta-analysis of observational data. However, these issues are not specific to the proposals in our paper, and we therefore do not consider them.

To reduce bias, prospectively planned meta-analyses are the ideal approach, but they are expensive, time-consuming and ignore the large amount of data already available. They also require model building with its attendant weaknesses.

Finally, an over-emphasis on bias results in questioning essentially all the results of observational studies and does not help in tackling medically and scientifically important questions.

6.4. Issues and limitations

Our third analysis option, studywise selected FP, may only be sensible when the sample size in each study is large and/or the effect of the factor of interest is strong. The reason is lack of power. In our example, the effect of age is weak, and this is reflected in very variable functions in each 'study' (see Figure 4), whereas the results for the strong predictor, nodes, are more plausible (see Figures 2 and 3). Further information on the relative merits of our three proposed options should be obtained from simulation studies.

An additional option, beyond the scope of the present paper, is a variant of our first option, overall FP2, in which the same power terms are used in all studies. It is based on related work on multivariate meta-analysis [42, 43]. Something similar was proposed by Rota and colleagues [28] for analysing published dose–response data. Instead of pointwise averaging of the individual functions, the FP2 parameter estimates are meta-analysed, using either fixed or random effects models. Comparisons with our approach will illustrate their similarities and differences.

It is obviously preferable where possible to use the same confounders (measured with the same techniques etc) in each study. In reality, however, the number of variables common to all studies may be small, raising the question of how or even whether to adjust. See Section 5.2 for further discussion of this issue.

Further general issues of meta-analysis are discussed in Section 5.

6.5. Summary

We propose the new strategy in the context of survival data analysed with Cox's proportional hazards model. None of the standard assumptions of the model must be appreciably violated in any of the studies. Transfer to other regression models is immediate. Our approach could substantially improve meta-analyses of continuous variables in observational studies. Because we combine three well-established steps (model selection to derive an index, selection of an FP function, averaging of functions) in a simple way, the method can easily be used and modified according to the requirements of a given meta-analysis. However, to provide better guidance, more experience with real data and the results of simulation studies are needed.

Appendix

The FP function selection procedure (FSP)

In step 2 we determine a functional form for each study by the fractional polynomial approach. We will consider 4 methods to determine power transformations. General power models were proposed by Box

and Tidwell [44]. Royston and Altman [11] formalised the simple power models and called them fractional polynomials of degree 1 (FP1), and extended them to FPs higher degree. Ignoring the intercept term (β_0), a polynomial of degree m in a single covariate Z may be written $\beta_1 Z^1 + \beta_2 Z^2 + \ldots + \beta_m Z^m$. In the generalisation to an FP function of degree m, written FP $m(z)$, the indices $1, \ldots, m$ are replaced with powers $p_1, \ldots, p_m$, giving $FPm(Z) = \beta_1 Z^{p_1} + \beta_2 Z^{p_2} + \ldots + \beta_m Z^{p_m}$.

As proposed by Royston and Altman [11], the powers $p_1, \ldots, p_m$ are chosen from a restricted set, $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3, \}$, where Z^0 denotes $\log Z$. The set includes no transformation ($p = 1$) and the reciprocal, logarithmic, square root and square transformations.

FP2 functions take the form $\beta_1 Z^{p_1} + \beta_2 Z^{p_2}$ or $\beta_1 Z^{p_1} + \beta_2 Z^{p_1} \log Z$ the latter being the so-called repeated-powers function obtained in the mathematical limit $p_2 \to p_1$.

In practice, the families of 8 FP1 and 36 FP2 functions (with an intercept term) provide a good fit to many biomedical datasets, and higher-order functions are rarely needed.

Ambler and Royston [45] and Sauerbrei and Royston [12, 46] give details of a function selection procedure (FSP).

1. Test the best FP2 model for z at the α level against the null model using 4 d.f.. If the test is not significant, stop, concluding that the effect of z is 'not significant' at the α level. Otherwise continue.
2. Test the best FP2 for z against a straight line at the α level using 3 d.f.. If the test is not significant, stop, the final model being a straight line. Otherwise continue.
3. Test the best FP2 for z against the best FP1 at the α level using 2 d.f.. If the test is not significant, the final model is FP1, otherwise the final model is FP2. End of procedure.

The test at step 1 is of overall association of the outcome with x. The test at step 2 examines the evidence for non-linearity. The test at step 3 is to choose between a simpler or more complex non-linear model.

Before applying the procedure, the user must decide on the nominal P $-$value (α) and on the degree (m) of the most complex FP model allowed. Typical choices will be $\alpha = 5\%$ and FP2 ($m = 2$). For FP1 function the modification is obvious, starting with a test of the best FP1 function against the null model on 2 d.f.

Acknowledgements

We thank Vincent Vinh-Hung for making the SEER data available in the format we used in our paper. We are grateful to the editor and three reviewers for detailed comments that helped us to strengthen the paper. Some of the work was carried out during Research-in-Pairs (RiP) visits to the Mathematisches Forschungsinstitut Oberwolfach, Germany, in 2008 and 2009. Royston was supported by UK Medical Research Council grant number MC_US_A737_0002.

References

1. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose–response data, with applications to meta-analysis. *American Journal of Epidemiology* 1992; **135**:1301–1309.
2. White IR. The level of alcohol consumption at which all-cause mortality is least. *Journal of Clinical Epidemiology* 1999; **52**:967–975.
3. Bagnardi V, Zambon A, Quatto P, Corrao G. Flexible meta-regression functions for modeling aggregate dose–response data, with an application to alochol and mortality. *American Journal of Epidemiology* 2004; **159**:1077–1086.
4. Sauerbrei W, Blettner M, Royston P. On alcohol consumption and all-cause mortality. *Journal of Clinical Epidemiology* 2001; **54**:537–538.
5. White IR. On alcohol consumption and all-cause mortality. response. *Journal of Clinical Epidemiology* 2001; **54**:538–540.
6. Butcher I, Maas AIR, Lu J, Marmarou A, Murray GD, Mushkadiani NO, McHugh GS, Steyerberg EW. Prognostic value of admission blood pressure in traumatic brain injury: results from the IMPACT study. *Journal of Neurotrauma* 2007; **24**:294–302.
7. Rovers MM, Glasziou P, Appelman CL, Burke P, McCormick DP, Damoiseaux RA, Little P, Saux NL, Hoes AW. Predictors of pain and/or fever at 3 to 7 days for children with acute otitis media not treated initially with antibiotics a meta-analysis of individual patient data. *Pediatrics* 2007; **119**:579–585.
8. Goebell PJ, Groshen S, Schmitz-Drager BJ, Sylvester R, Kogevinas M, Malats N, Sauter G, Barton GH, Waldman F, Cote RJ. The International Bladder Cancer Bank: proposal for a new study concept. *Urological Oncology* 2004; **22**:277–284.
9. Royston P, Parmar MKB, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Statistics in Medicine* 2004; **23**:907–926.
10. Sutton AJ, Higgins JPT. Recent developments in meta-analysis. *Statistics in Medicine* 2008; **27**:625–650.

11. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics* 1994; **43**(3):429–467.
12. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society, Series A* 1999; **162**:71–94.
13. Royston P, Sauerbrei W. *Multivariable model-building. A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. John Wiley & Sons: Chichester, 2008.
14. Tai P, Cserni G, van de Steene J, Vlastos G, Voordeckers M, Royce M, Lee SJ, Vinh-Hung V, Storme G. Modeling the effect of age in T1–2 breast cancer using the SEER database. *BMC Cancer* 2005; **5**:130.
15. Sauerbrei W, Royston P, Bojar H, Schmoor C, Schumacher M, The German Breast Cancer Study Group. Modelling the effects of standard prognostic factors in node positive breast cancer. *British Journal of Cancer* 1999; **79**:1752–1760.
16. Vinh-Hung V, Burzykowski T, Cserni G, Voordeckers M, van de Steene J, Storme G. Functional form of the effect of the numbers of axillary nodes on survival in early breast cancer. *International Journal of Oncology* 2003; **22**:697–704.
17. Trivella M, Pezzella F, Pastorino U, Harris AL, Altman DG, The Prognosis in Lung Cancer (PILC) Collaborative Study Group. Microvessel density as a prognostic factor in non-small-cell lung carcinoma: a meta-analysis of individual patient data. *Lancet Oncology* 2007; **8**:488–499.
18. Look MP, van Putten WLJ, Duffy MJ, Harbeck N, Jarle I, Christensen CT, Thomssen C, Kates R, Spyratos F, Fernö M. *et al.* Pooled analysis of prognostic impact of urokinase-type plasminogen activator and its inhibitor PAI-1 in 8377 breast cancer patients. *Journal of the National Cancer Institute* 2002; **94**:116–128.
19. Royston P, Sauerbrei W, Look MP. *Standardizing continuous markers across studies*, 2010. Submitted.
20. Riley RD, Sauerbrei W, Altman DG. Prognostic markers in cancer: the evolution of evidence from single studies to meta-analysis, and beyond. *British Journal of Cancer* 2009; **100**:1219–1229.
21. Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *International Journal of Epidemiology* 1999; **28**:1–9.
22. Smith-Warner SA, Spiegelman D, Ritz J, Albanes D, Beeson WL, Bernstein L, Berrino F, van den Brandt PA, Buring JE, Cho E. *et al.* Methods for pooling results of epidemiologic studies: the Pooling Project of Prospective Studies of Diet and Cancer. *American Journal of Epidemiology* 2006; **163**:1053–1064.
23. Thompson S, Kaptoge S, White I, Wood A, Perry P, Danesh J. the Emerging Risk Factors Collaboration. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. *International Journal of Epidemiology* 2010; **39**(5):1345–1359.
24. Sharp S, Sterne J. Meta analysis. *Stata Technical Bulletin* 1997; **38**:9–14.
25. Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. Cochrane Working Group. *Statistics in Medicine* 1995; **14**:2057–2079.
26. Altman DG, Trivella M, Pezzella F, Harris AL, Pastorino U. *Systematic review of multiple studies of prognosis: the feasibility of obtaining individual patient data. I. Advances in Statistical Methods for the Health Sciences*. Birkhäuser: Boston, 2006; 3–18.
27. Dickersin K. Systematic reviews in epidemiology: why are we so far behind?. *International Journal of Epidemiology* 2002; **31**:6–12.
28. Rota M, Belloco R, Scotti L, Tramacere I, Jenab M, Corrao G, Vecchia CL, Boffetta P, Bagnardi V. Random-effects meta-regression models for studying dose–response relationship, with an application to alcohol and esophageal squamous cell carcinoma. *Statistics in Medicine* 2010; **29**:2679–2687.
29. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* 1999; **28**:964–974.
30. MacCallum R, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. *Psychological Methods* 2002; **7**:19–40.
31. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* 2006; **25**:127–141.
32. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52 705 women with breast cancer and 108 411 women without breast cancer. *Lancet* 1997; **350**:1047–1059.
33. Sauerbrei W, Holländer N, Riley RD, Altman DG. Evidence-based assessment and application of prognostic markers: the long way from single studies to meta-analysis. *Communications in Statistics* 2006; **35**:1333–1342.
34. Boffetta P, Soutar A, Cherrie JW, Granath F, Andersen A, Anttila A, Blettner M, Gaborieau V, Klug SJ, Langard S. *et al.* Mortality among workers employed in the titanium dioxide production industry in Europe. *Cancer Causes and Control* 2004; **15**:697–706.
35. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Research and Treatment* 1992; **22**:207–219.
36. Kattan MW. Judging new markers by their ability to improve predictive accuracy. *Journal of the National Cancer Institute* 2003; **4953**:634–635.
37. Royston P, Sauerbrei W. Improving the robustness of fractional polynomial models by preliminary covariate transformation. *Computational Statistics and Data Analysis* 2007; **51**:4240–4253.
38. Draper D. Assessment and propagation of model selection uncertainty (with) discussion. *Journal of the Royal Statistical Society: Series B* 1995; **57**:45–97.
39. Faes C, Aerts M, Geys H, Molenberghs G. Model averaging using fractional polynomials to estimate a safe level of exposure. *Risk Analysis* 2007; **27**:111–123.
40. Greenland S. Multiple-bias modelling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society (Series A)* 2005; **168**:267–306.
41. Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society (Series A)* 2009; **172**:21–47.

42. Jackson D, White IR, Thompson SG. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine* 2010; **29**:1282–1297.
43. White IR. Multivariable random-effects meta-analysis. *Stata Journal* 2009; **9**(1):40–56.
44. Box GEP, Tidwell PW. Transformation of the independent variables. *Technometrics* 1962; **4**:531–550.
45. Ambler G, Royston P. Fractional polynomial model selection procedures investigation of Type I error rate. *Journal of Statistical Computation and Simulation* 2001; **69**:89–108.
46. Sauerbrei W, Royston P. Corrigendum: Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society, Series A* 2002; **165**:399–400.