Stockholms universitet

A point-wise average approach in dose-response meta-analysis of summarized data for binary outcomes

Ilias Thomas

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2015:5
Matematisk statistik
Maj 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Master Thesis **2015:5**
http://www.math.su.se

A point-wise average approach in dose-response meta-analysis of summarized data for binary outcomes

Ilias Thomas*

May 2015

Abstract

In this thesis we propose a point-wise averaging approach for dose-response meta-analysis of aggregated data. Relative to the common approach of averaging regression coefficients the method proposed allows for more flexibility. Recently proposed, the point-wise approach is a new strategy to perform meta-analysis on individual patient data, but has not been investigated in the context of aggregated data. Each individual study is allowed to follow a different dose-response trend using predictor transformations such as splines or fractional polynomials. The predicted outcomes are then averaged across studies at specific values of the quantitative predictor. The methodology is described in detail and is applied to survival data from 9 Registries of the Surveillance, Epidemiology, and End Results Pro- gram (SEER) of the United States, involving breast cancer patients. The performance of the method is evaluated against the dose-response meta-analysis of individual patient data analysis. The method has been tested using simulated studies of common dose-response relations (i.e. linear, U-shaped, J-shaped). Overall, the point-wise approach on aggregated data produces similar results in comparison with the same analysis on individual patient data and comparable results with the true underlying shapes of simulated studies.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: elias_thomas_1987@hotmail.com . Supervisor: Taras Bodnar.

Acknowledgements

First of all, I want to express my gratitude to Dr. Nicola Orsini at Karolinska institute for giving me the opportunity to participate in this project working with him at the department of Environmental Medicine. I would also like to thank associate Prof. Taras Bodnar for his supervision and advise throughout the writing process of the thesis. Special thanks to Alessio and Andrea for the useful conversations and advises. To my family for their unconditional love and support and to Eleni for always being there for me. Finally to all my friends for making me feel at home away from home.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Epidemiologists, clinical investigators and those working on health policy are particularly interested in quantifying the observed exposure-outcome relationship of disease onsets or other events, and in particular how the risk of the outcome changes across the different levels of exposure. The term exposure is not limited to a single definition, but instead is used to describe different varieties of dose e.g. could be a treatment levels on a randomized trial, the amount of intake of a certain nutrient or the actual exposure to an environmental factor.

In recent years, as the number of studies concerning dose-response relations increases, meta-analysis has become more popular and the number of publications applying dose-response meta-analysis is rising exponentially year by year as can be seen in figure 1.1

On a survey we conducted on the Web of Science the number of publications concerning dose-response meta-analysis rose from 30 publications in 2004 to 222 publications in 2014. Of the 222 publications in 2014, 140 qualified for our survey, where meta-analysis was performed both on individual patient data (IPD) and aggregated data. About 5 % (8) of those publications concerned IPD and only 37,5 % (3) of those considered nonlinear trends in the dose-response relation whether the other 62,5 % (5) did not present a plot for the dose-response shape. Furthermore about 95 % (132) of the publications performed meta-analysis on aggregated data and

out of those, 40 % (53) didn't present the dose-response shape, while 56 % (74) presented or considered a nonlinear trend, with only 4 % (5) assuming a linear trend. The most popular fields where dose-response meta-analysis is used are oncology, public environmental occupational health, nutrition dietetics, endocrinology metabolism and general internal medicine.

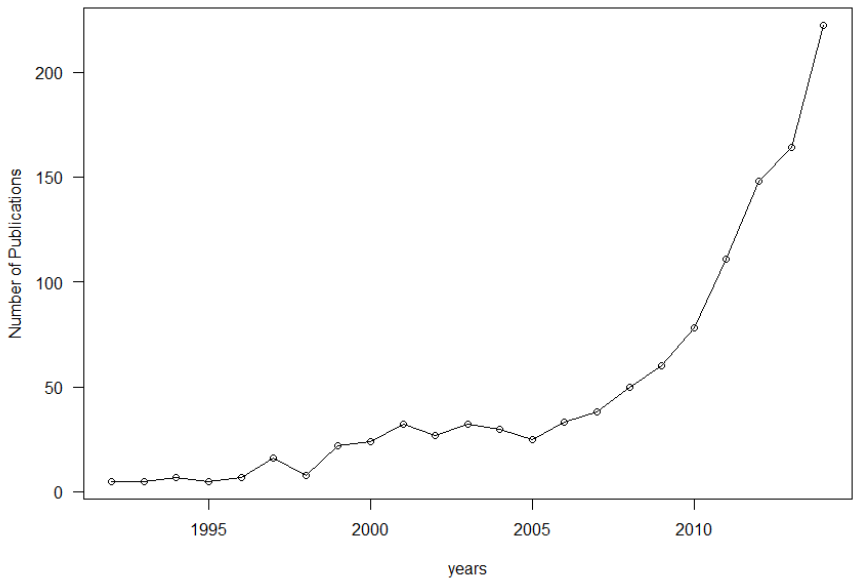

Figure 1.1: The publication of papers concerning dose-response meta-analysis over the years

Despite the increasing popularity of dose-response meta-analysis, the methods used are not evolving accordingly. Since the method first described by Greenland and Longnecker [1], few methodological articles have been published on how to perform meta-analysis for non-linear dose-response shapes, the most notable including: Bagnardi et al [2]; Orsini et al. [3]; Rota et al. [4]; Sauerbrei and Royston [5]; Shi and Copas [6]; Liu et al [7]; Gasparini et al [8] Thomas et al [9].

The aim of this thesis is to fill a gap in the literature, that is to introduce a new more robust and flexible approach to perform dose-response meta-analysis of aggregated data for binary outcomes. Originally proposed by

Sauerbrei and Royston [5], the point-wise approach is a new strategy to perform meta-analysis on IPD for continuous predictors. This approach, however, has not been investigated in the context of aggregated data. The need of more efficient ways to perform meta-analysis on summarized studies arises also from the fact that in a large majority (86%) of the cases were we observe continuous risk factors, the investigators chose to categorise the exposure (Turner et al. [10]).

In this novel approach, instead of averaging the regression coefficients of a common pre-specified model, we apply a point-wise averaging procedure of study-specific trends. Each individual study is allowed to have a different dose-response shape and predicted outcomes are then pooled at specific values of the quantitative predictor.

In the following sections of the thesis we will outline and discuss the proposed approach. The methodology will be implemented, using data from 9 registries of the Surveillance, Epidemiology and, End Results program of the United States, concerning data of individual breast cancer patients from different population-based cancer studies. A direct comparison of the point-wise approach for individual patient data and the point-wise method for summarized data is made, to point-out similarities and differences. The point-wise approach is then evaluated in simulated studies of the most common dose-response relations. Finally, a short discussion is made.

Chapter 2

Methods

The most common approach for trend estimation consists of quantifying the dose-response relation and examining the change of the risk across exposure levels. Summarized data are usually reported as a series of the dose categories, with their corresponding relative risks, with one of the categories serving as reference. Here the term relative risk will be used as a generic term for risk ratio, hazard ratio, rate ratio or odds ratio. In tables 2.1,2.2, and 2.3, examples of aggregated data are shown in the case of a cumulative incidence,incidence-rate, or case-control study.

Table 2.1: Example of aggregated data of a cumulative incidence study

Category	**Dose**	**Cases**	**Total Patients**	**Relative Risk**	**95 % CI**
Category 0	x_0	A_0	N_0	1	(1, 1)
...	...	...	...	....	(..., ...)
...	...	...	...	...	(..., ...)
Category T	x_T	A_T	N_T	$RR_T = e^{y_T}$	(lb_T, ub_T)

Table 2.2: Example of aggregated data of an incidence-rate study

Category	**Dose**	**Cases**	**Person-time**	**Rate Ratio**	**95 % CI**
Category 0	x_0	A_0	N_0	1	(1, 1)
...	...	...	...	....	(..., ...)
...	...	...	...	...	(..., ...)
Category T	x_T	A_T	N_T	$RR_T = e^{y_T}$	(lb_T, ub_T)

Table 2.3: Example of aggregated data of a case-control study

Category	**Dose**	**Cases**	**Controls**	**Odds Ratio**	**95 % CI**
Category 0	x_0	A_0	B_0	1	(1, 1)
...	...	...	...	....	(..., ...)
...	...	...	...	...	(..., ...)
Category T	x_T	A_T	B_T	$RR_T = e^{y_T}$	(lb_T, ub_T)

2.1 Individual trend estimation for the studies

In order to analyse the dose-response relation within each study a log-linear regression model can be used. Let us first consider a single study, which we index with i. We can express the model as follows:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \tag{2.1}$$

where $\mathbf{y}_i$ is a $T_i \times 1$ vector of reported log-relative risks (not including the reference one), $\mathbf{X}_i$ is a $T_i \times p$ matrix of non-stochastic covariates containing the value of exposure and/or some other transformations of it, $\boldsymbol{\beta}_i$ is a $1 \times p$ vector of unknown regression coeffcients for the i-th study and $\boldsymbol{\epsilon}_i$ is a $T_i \times 1$

vector that expresses the random errors. The variance-covariance matrix $Cov(\epsilon_i) = E(\epsilon_i \epsilon_i')$ is equal to the following symmetric matrix

$$Cov(\epsilon_i) = \mathbf{\Sigma}_i = \begin{bmatrix} \sigma_{11}^2 & & & & \\ . & \ddots & & & \\ . & & & & \\ \sigma_{t1}^2 & & \sigma_{tt}^2 & & \\ . & & & & \\ . & & & \ddots & \\ \sigma_{T_i1}^2 & \dots & \sigma_{T_it}^2 & \dots & \sigma_{T_iT_i}^2 \end{bmatrix}$$

The model in equation 2.1 has no intercept, assuming that the exposure variable has value 0 at the reference category and the log relative risk at the reference category is set to 0 (the relative risk is 1). If we consider an exposure with non-zero reference category, such as BMI or energy intake, we can center the levels of exposure such that the reference value becomes zero. The goal is to estimate the coefficients $\boldsymbol{\beta}_i$ for the log-linear model i.e. the change in the natural logarithm of the relative risk per unit of exposure within each study. The $\boldsymbol{\beta}_i$'s are estimated using generalized least squares regression as proposed by Greenland and Longnecker [1].

2.2 Generalized least squares regression

Under the assumption that the variance-covariance matrix is known the trend or regression coefficients can be efficiently estimated using the generalized least squares regression. Referring to $\mathbf{\Sigma}_i$, the covariance matrix between the log-relative risks, the method requires minimizing $(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i)')(\mathbf{\Sigma}_i^{-1})(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i)$ with respect to $\boldsymbol{\beta}_i$. The estimator $\hat{\boldsymbol{\beta}}_i$ of the trend coefficients β_i is finally:

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i)^{-1} \mathbf{X}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{y}_i \tag{2.2}$$

where the estimated covariance matrix of β_i, V_i is

$$\mathbf{V}_i = (\mathbf{X}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i)^{-1} \tag{2.3}$$

2.3 Estimating the covariance using the Greenland and Longnecker method

As mentioned before, published dose-response data are typically reported as a series of dose specific relative risks, with one category serving as the common referent group. Therefore, the elements of $\mathbf{y}_i$ are not independent and the off-diagonal elements of $\boldsymbol{\Sigma}_i$ are not zero. This section describes the method and formulas needed to estimate all the elements of $\boldsymbol{\Sigma}_i$.

If we let A_t be the number of cases at each exposure level t; B_t, the number of controls (for case-control data) at each exposure level t; and N_t, the total number of subjects (for cumulative incidence data) or the total person-time (for incidence rate data) in exposure level t,where the range of t is 0 (referent group) to T (the number of non-reference exposure levels). Examples of data from case-control studies, cohort studies (incidence rate data) and for cumulative incidence studies can be found in tables 2.1,2.2 and 2.3.

The Greenland-Longnecker method to estimate the elements of the covariance matrix assumes that the correlation between the crude log(RR) risks are approximately equal to the correlation of the adjusted log(RR). The method to estimate efficient point estimators and consistent variance estimators for a series of reported log-relative risks comprises of the following steps:

1. Given the adjusted log(RR) and the total number of cases and non-cases at each exposure level t of each study, solve for the fitted cell counts in each data table. This method requires a simple fitting algorithm based on the Newton method.

2. The next step is to approximate the correlation among the log(RR) for t other than l, as $s_{tl} = s_0/(s_t s_l)^{1/2}$, where s_0 is the common covariance while s_t and s_l are the variances of the log(RR). The equations to calculate the covariances and the variances differ from study to study. For a case-control study we have: $s_0 = (\frac{1}{A_0} + \frac{1}{B_0})$ and $s_t = (\frac{1}{A_t} + \frac{1}{B_t} + \frac{1}{A_0} + \frac{1}{B_0})$, for incidence-rate studies $s_0 = (\frac{1}{A_0})$ and $s_t = (\frac{1}{A_0} + \frac{1}{A_t})$ and for cumulative incidence studies $s_0 = (\frac{1}{A_0} - \frac{1}{N_0})$ and $s_t = (\frac{1}{A_t} - \frac{1}{N_t} + \frac{1}{A_0} - \frac{1}{N_0})$.

3. Approximate the multivariate asymptotic covariances between the adjusted log relative risks as $\sigma_{tl} = r_{tl} \times (\sigma_t \sigma_l)^{1/2}$, where r_{tl} represent the correlations estimated in the previous step while σ_t and σ_l are the variances of the adjusted log relative risks, (the diagonal elements of matrix $\mathbf{\Sigma}_i$), defined as the length of the confidence interval on the log scale divided by the square of the (1-$\alpha/2$) - level standard normal deviate for all t different than l. This method relies on the assumption that the correlation matrix of the unadjusted and adjusted relative risks are nearly the same. In other terms, it is valid when there is no confounding (or at least slight) by other model covariates in the published results of each study included in the meta-analysis.

2.4 Predictions for each study using the log-linear model

Based on the trend coefficients extracted from the individual $i = 1, ..., I$ studies, we use the log-linear model in equation 2.1 in order to make predictions for the response variable $\hat{\mathbf{y}}_i = \mathbf{X}_i \hat{\boldsymbol{\beta}}_i$ (predicted outcomes), given the desired range $j = 1, ..., n$ of the exposure values. If the exposure is modeled using transformations to allow for a non-linear relation of the dose-response curve

(Crippa and Orsini [11]; Bagnardi and al. [2]), $\hat{\boldsymbol{\beta}}_i$ would be a vector of $\mathbf{s} = 1, ..., p$ coefficients, the length of which depends on the choice of transformation (splines, polynomials), and $\mathbf{X}$ would be a $n \times p$ matrix, where the first column identifies the exposure variable and the other p-1 columns could represent transformations of the exposure, based on the modeling of $\hat{\boldsymbol{\beta}}_i$ (Orsini et al., [12]; Berlin et al. [13])

$$\mathbf{X} = \begin{bmatrix} x_{11} & ... & x_{1s} & ... & x_{1p} \\ . & . & & & . \\ . & . & & . & \\ x_{j1} & ... & x_{js} & ... & x_{jp} \\ . & . & & & . \\ . & . & & & . \\ x_{n1} & ... & x_{ns} & ... & x_{np} \end{bmatrix}$$

Using equation 2.1 for each study, together with the vector $\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{i1}, .., \hat{\beta}_{ip})$ of the estimated trend coefficients and $\mathbf{X}$ we acquire the study-specific estimated log-relative risks.The corresponding standard errors of the predicted log relative risks are calculated as

$$\mathbf{s}_i = \sqrt{diag(\mathbf{X} \times cov(\hat{\boldsymbol{\beta}}_i) \times \mathbf{X}')} \tag{2.4}$$

From the performed procedure, two vectors of values are obtained for each study, one with the predicted outcomes (log relative risks at each level of the exposure) and one with the corresponding standard errors.

2.5 Point-wise dose-response meta-analysis

Having acquired the predicted outcomes for each study for the range of the exposure, the final step of the approach is to perform the point-wise meta-analysis. Two matrices, $\hat{\mathbf{Y}}$ and $\mathbf{S}$ with $n \times I$ dimensions can be constructed, using as rows the vectors described above

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{y}_{11} & \dots & \hat{y}_{1j} & \dots & \hat{y}_{1n} \\ . & . & & & . \\ . & . & & . & \\ \hat{y}_{i1} & \dots & \hat{y}_{ij} & \dots & \hat{y}_{in} \\ . & . & & & . \\ . & . & & & . \\ \hat{y}_{I1} & \dots & \hat{y}_{Ij} & \dots & \hat{y}_{In} \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} s_{11} & \dots & s_{1j} & \dots & s_{1n} \\ . & . & & & . \\ . & . & & & . \\ s_{i1} & \dots & s_{ij} & \dots & s_{in} \\ . & . & & & . \\ . & . & & & . \\ s_{I1} & \dots & s_{Ij} & \dots & s_{In} \end{bmatrix}$$

where $\hat{y}_{ij}$ is the predicted outcomes of i study at each level j of the exposure and s_{ij} the respective standard errors. The columns of the $\hat{\mathbf{Y}}$ matrix each correspond at a different level of the exposure and represent the respective log relative risk estimates of the different studies acquired at that level, I different estimates, for I number of studies.

Treating the values of the predicted outcomes at each level of the quantitative predictor (the columns of the $\hat{\mathbf{Y}}$ matrix) as a distribution about a certain mean, it is our intention to estimate that mean.

Different weighs are assigned to each study and the combined effect of the studies at a specific level of the exposure is given as the weighed average across the different studies, the weights at each level being computed as the inverse-variance of each estimate at that specific level of the exposure as described by Borenstein and al. (2010) [14].

With this application at each column of $\hat{\mathbf{Y}}$, the matrix will be reduced into a vector of values $\mathbf{M} = (M_1, ..., M_J, .., M_n)$, j=1,..,n for the levels of the exposure, that yields the dose-response relation, with each element of $\mathbf{M}$

representing the estimated log-relative risk of the outcome, compared to a reference value of the exposure. The elements of the vector $\mathbf{M}$ for the I number of different studies are computed as

$$M_j = \frac{\sum_{i=1}^{I} W_{ij}\hat{y}_{ij}}{\sum_{i=1}^{I} W_{ij}} \tag{2.5}$$

Where $\hat{y}_{ij}$ is the different effect sizes of the I studies at the value j of the exposure and W_{ij} is the respective weight assigned to each of the I studies at the value j of the exposure. The weights are calculated as

$$W_{ij} = \frac{1}{V_{ij} + T_j^2} \tag{2.6}$$

Where V_{ij} is the variance of each study I at the level j of the exposure (calculated from the standard error s_{ij}) and T_j is the between study variance at each point j of the exposure, which is common to all studies. The formula in equation 2.6 considers a random effects model, whether in the case of a fixed effects model, T_j^2 would be 0. For each of the estimates of the $\mathbf{M}$ vector the corresponding standard error is calculated, creating a vector of standard errors $\mathbf{SE}_M = (SE_{M_1}, .., SE_{M_j}, .., SE_{M_n})$ analogous to $\mathbf{M}$, where

$$SE_{M_j} = \frac{1}{\sqrt{\sum_{i=1}^{k} W_{ji}}} \tag{2.7}$$

so that the lower and upper bounds of the point-wise confidence intervals can be constructed as

$$CI_{M_j} = (M_j - Z_{a/2}SE_{M_j}, M_j - Z_{a/2}SE_{M_j}) \tag{2.8}$$

where a is derived from the confidence level of the intervals and is 0.05 if we choose 95 % confidence intervals. The Z value follows the cumulative normal distribution function and $Z_{a/2}$ would be 1.96 in the case where a equals 0.05

2.6 Point-wise heterogeneity

One of the main issues in meta-analysis, is addressing the question, if there is a significant statistical heterogeneity between the studies and to quantify the level of heterogeneity. The Q statistic at the j-th exposure level can be defined as

$$Q_j = \sum_{i}^{I} W_{ij}(Y_{ij} - M_i)^2 \tag{2.9}$$

derived here to fit the point-wise approach and measure the Q value at each point of the exposure level. Under the null hypothesis, Q has approximately a χ^2 distribution with $I - 1$ degrees of freedom where I is the total number of studies. If the p-value derived from this statistic is small, we may infer that there is some problem with the model; e.g., perhaps statistical heterogeneity is present or there is some unaccounted-for bias. If, however, the p-value is large, we can conclude only that the test did not detect any significant heterogeneity in the analysis, not that there is no heterogeneity at all. Q statistic (like most fit statistics) has low power; i.e., its sensitivity to model problems is limited. Higgings and Thompson [15], derived a statistic to measure heterogeneity independently of the number of studies. The I^2 statistic is defined as

$$I_j^2 = max\{0, (Q_j - df)/Q_j\} \tag{2.10}$$

that can be used to calculate the percentage of the variance of the study estimates explained by heterogeneity between the studies.
Based on the approach we are proposing, we can extend the concept of heterogeneity between studies, not only to a single statistic, but to a vector of statistics $\mathbf{I}^2 = (I_1^2, .., I_j^2, .., I_n^2)$ for each point of the quantitative predictor. One can inspect the heterogeneity across the range of the exposure and identify patterns, for example if the heterogeneity increases or decreases at a certain levels of the exposure.

2.7 Background in fractional polynomials and spline regression models

In this section we briefly discuss the use of Fractional Polynomial and Spline regression to model the exposure

2.7.1 Fractional polynomial regression

The fractional polynomials of order P are expressed as a family of model functions of the single continuous and positive covariate x, in our case the exposure. The model

$$y = \sum_{p=1}^{P} \beta_p x^{p_p} \tag{2.11}$$

represents log(RR) or the relative risk on the normal scale, if we choose so, as a function of power transformations of the exposure variable x, where β_p is the vector of coefficients as described before and p_p are the elements of vector p, the power vector of the transformation. As stated by Bagnardi and al. [2], it is efficient to only consider the family of second order polynomials, to allow for non-monotonic curves of the dose-response relation and also

because the second order families offer great flexibility. More precisely, if we let p_1 and p_2 belong to the set of values P=(-2,-1,-0.5,0,0.5,1,2,3), the model

$$y = \begin{cases} \beta_1 x^{p_1} + \beta_2 x^{p_2} & \text{if } p_1 \neq p_2, \\ \beta_1 x^p + \beta_2 (x^p log(x)) & \text{if } \mathrm{p}_1 = p_2 = p. \end{cases} \tag{2.12}$$

can account for a prosperous set of dose-response shapes, including the likes of J-shape, U-shape and most positive and negative dose-response associations. Using the Box-Tidwell transformation, in the case where $p_i = 0$, x^{p_i} reduces to log(x) and remains x^{p_i} otherwise. For all the possible combinations from the set of P, 36 models can be considered in total. The model with the best fit across the models generated, can be defined as the one with the highest likelihood or lowest deviance. In the analysis performed in this thesis, we choose the models minimizing the Akaike information criterion (AIC), as suggested by Rota et al. [4]. Broadly, non-nested models may be compared using this statistic. Letting L be the maximized value of the likelihood function and k the number of estimated parameters in the the model the AIC value is defined as: $AIC = k - 2log(L)$

2.7.2 Spline regression

A spline function is a smoothly piecewise polynomial of order q. The observed range of the exposure is categorized based on a number of K knots and a spline model is fitted within the categories corespondent to the number of knots. The cubic spline models have the most common use in the literature, because they offer great flexibility for data fit. Dividing the exposure into categories, a third order polynomial can be fitted in each category and the model

$$y = \sum_{z=1}^{3} \beta_{0z} x^z \tag{2.13}$$

can be used to describe the dose-response association in a single category. One key assumption for the model in equation 2.13 is that the function is 2 times continuously differentiable in the range of the exposure. Fitting the model in all categories of the exposure would lead to the single function of the scalar variable x

$$y = \sum_{z=1}^{3} \beta_{0z} x^z + \sum_{k=1}^{K} \beta_{k3} (x - l_k)^3 \tag{2.14}$$

where l_k are the positions of the knots $k = 1, .., K$. In order to avoid odd behaviour of the fitted curve in the tails, the cubic spline can be restricted to be linear there, leading to the restricted cubic spline regression model

$$y = \beta_{01} x + \sum_{k=1}^{K-2} \beta_{k3} [(x - l_k)^3 - \frac{(x - l_{K-1})^3 (l_K - l_k)}{l_K - l_{K-1}} + \frac{(x - l_K)^3 (l_{K-1} - l_k)}{l_K - l_{K-1}}] \tag{2.15}$$

as described by Durrleman and Simon (1989) [16] and Desquilbet and Mariotti (2010) [17]. That model would yield K-1 coefficients describing the exposure. For notation reasons we can describe the model in equation 2.15,when modelling with three knots as

$$y = \beta_1 x + \beta_2 f(x) \tag{2.16}$$

which is a model with two coefficients and a function containing the tranformation of the exposure as seen in equation 2.15. A key issue with spline regression is the selection of the number and the position of the knots in the range of the exposure. Since we force the function to be linear in the tails, we choose the position of those knots not to be too far from the extremes, depending on the number of total knots. Choosing predefined knots at fixed percentiles of the exposure's distribution, is a good approach as it ensures

that enough points are available in each interval. Recommended equally spaced quartiles can be found on Harrell [18],although the location of the knots is not very important in practice which is one of the main advantages when working with splines.

2.8 Cox regression

In the analysis performed in later chapters, the Cox regression model is used to model the survival data. A key reason for the popularity of the Cox model is that good estimates of regression coefficients and hazards ratios of interest can be obtained for a wide variety of data situations. Furthermore, the Cox model is preferred over the logistic model when survival time is available and there is censoring. That means that the Cox model uses more information that the logistic model.

2.8.1 The formula for the Cox model

The standard notation for the Cox regression model is given in the equation:

$$h(t, \mathbf{X}) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i} \tag{2.17}$$

where the vector $\mathbf{X} = (X_1, ..., X_p)$ contains the explanatory variables of the model. The formula implies that the hazard at time t is a product of two quantities. The first one, $h_0(t)$ is called the baseline hazard function, because if all the explanatory variables become 0, then the formula would reduce to $h_0(t)$, in other words the baseline hazard. This quantity is non-parametric. The second quantity of equation 2.17 is an exponential expression e to the linear sum of $\beta_i X_i$ over the explanatory variables, which is fully parametric and the is the reason that the Cox model is called semi-parametric.

2.8.2 Maximum likelihood estimators for the Cox PH model

As in usual regression models the ML estimates of the cox model parameters are derived by maximizing the a likelihood function, denoted as L. The likelihood function is a mathematical expression which describes the joint probability of obtaining the data observed on the subjects in the study as a function of the unknown parameters (the coefficients) in the model being considered. The formula for the Cox model is actually called a partial likelihood function, because it considers probabilities only for the subjects that fail, and does not explicitly considers probabilities for the subjects that are censored.

Once the likelihood function is formed for a given model, the next step would be to maximize the function. An appropriate way to do so is to to solve

$$\frac{lnL}{d\beta_i} = 0 \tag{2.18}$$

for all i = 1,...,p the number of the parameters

2.8.3 Hazard ratio

The hazard ratio is defined as the hazard for one individual over the hazard for a different individual. The two individuals compared are usually distinguished by their values for the set of the predictors, that is the vector $\mathbf{X}$. If we denote the set of predictors for an individual as $\mathbf{X}$ and the set of predictors for a different individual as $\mathbf{X}^*$, the expression for the hazard ratio is

$$\hat{HR} = \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \frac{\hat{h}_0(t) e^{\Sigma_{i=1}^{p} \hat{\beta}_i X_i^*}}{\hat{h}_0(t) e^{\Sigma_{i=1}^{p} \hat{\beta}_i X_i}} = e^{\Sigma_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)} \tag{2.19}$$

2.8.4 The proportional hazards assumption

One of the most basic assumptions of the Cox regression model is the one of the proportional hazards, which requires that they are constant over time, or equivalently, that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time. That would mean the expression 2.19 would reduce to

$$\frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \hat{\theta} \tag{2.20}$$

where $\hat{\theta}$ is a constant over time. The variables for which the PH assumption holds are called time-independent variables whether the variables for which the PH assumption does not hold are called time-dependent variables. If time dependent variables are considered then a variant of the Cox regression model might still be used, but such a model no longer satisfies the PH assumption, and is called the extended Cox model.

Chapter 3

Results

3.1 Individual patient data

We illustrate the point-wise average approach here using survival data of breast cancer patients from 9 Registries of the Surveillance, Epidemiology, and End Results Program (SEER) of the United States (http://seer.cancer.gov), which contains individual patient data from different population-based cancer studies. The dataset contains information about 712.319 breast cancer patients from nine registries: San Francisco-Oakland, Connecticut, Metropolitan Detroit, Hawaii, Iowa, New Mexico, Seattle (Puget Sound), Utah and Metropolitan Atlanta, which will be treated as different studies. Based on suggestions from Hung and al. [19] and Tai et al. [20] for the same dataset, we selected only the female patients, without previous history of cancer.

Additional criteria for the selection of the subjects were: the patients had cancer directed surgery where the type of surgery was either partial/less than total mastectomy or modified radical/total (simple) mastectomy. Radiotherapy status was known and reported as delivered or not delivered after the surgery. The invasive type of carcinoma was histologically confirmed. The patient's size of the primary tumor was known and was smaller than 50 mm. The laterality, was specified on the reports as either right or left origin of primary. At each of the surgeries, axillary dissection had been performed, with known number of nodes examined and number of positive nodes. For

all the patients, no known internal mammary node(s) was involved and there were no distant metastases.

Subjects were excluded, when their race or the month of their diagnosis were unknown and when the reporting source was other that a hospital. Subjects were also excluded when their reported number of positive nodes surpassed the number of examined nodes.

The only event of interest is defined as death from breast cancer before the cut-off date of 31 December 1999. After refining the data with the criteria described above, there were a total of 84.404 patients that qualified, with 8520 total events. The distribution of the patients within each study is shown in table 3.1.

Table 3.1: Number of individuals and number of events ,of each study with the median follow up time

Study	**Individuals**	**Events**	**(%)**	**Follow-up**
San Francisco-Oakland	14270	1399	9,8	6
Connecticut	11863	1177	9,9	6
Metropolitan Detroit	13939	1603	11,5	6
Hawaii	3895	291	7,4	6
Iowa	12409	1296	10,4	6
New Mexico	4705	470	9,9	6
Seattle (Puget Sound)	13474	1166	8,7	6
Utah	3118	334	10,7	6
Metropolitan Atlanta	6731	784	11,6	6

To highlight the approach, in connection with chapter 2, in the analyses we perform, we treat the death of breast cancer as the outcome of interest, and the number of positive nodes found in the patients as the exposure. For

all the analyses, the statistical program R was used, with the implementation of the packages *survival*, *dosresmeta*, *metafor* and *rms*.

3.1.1 Description of the Cox model

We used a Cox regression model to estimate the log hazard rations of the dose-response relation. Although, we treated the number of positive nodes as the exposure of interest, but also adjusted for other factors: age, race, marital status, histological confirmed cancer, grade of the cancer, tumor size, radiation sequence after the surgery and the presence or not of estrogen or progesterone receptors, following the suggestions of Sauerbrei and Royston [5]. The final model was:

$$h(t, X_1, \mathbf{C}_1, .., \mathbf{C}_9) = h_0(t)e^{(\beta_1 f_1(X_1)+\beta_2 f_2(X_1)+\sum_{i=1}^{9}\boldsymbol{\beta}_{C_i}\mathbf{C}_i)} \tag{3.1}$$

where the exposure of interest X_1 is transformed using two functions either with restricted cubic splines or second order fractional polynomials, as described in chapter 2.7. In the case of restricted cubic splines $f_1(X_1)$ will reduce to X_1. The rest of the explanatory variables in the model 2.18, noted as $\mathbf{C}_i$ represent the confounding factors of the model and could for each adjusting factor have length of 1 (linear or binary variable), 2 (quadratic form of a continuous variable) or 3 (variable with 4 categories), with their corresponding coefficients. The $\mathbf{C}_i's$ are described as follows:

- $\mathbf{C}_1$: Whether or not the invasive carcinoma was histologicaly confirmed, positive or negative.
- $\mathbf{C}_2$: Whether or not the subjects where married.
- $\mathbf{C}_3$: The grade of the tumor. Could vary between 1,2 or 3. For this category, missing values were allowed and they were coded as a different category.

- $\mathbf{C}_4$: The age of the patient at diagnosis. It was given a quadratic form.
- $\mathbf{C}_5$: Endocrine receptor positive or negative. They grow in response to the hormone estrogen.
- $\mathbf{C}_6$: Endocrine receptor positive or negative. They grow in response to another hormone, progesterone
- $\mathbf{C}_7$: The radiation sequence before or after the surgery, treatment or no treatment. It was forced into the model, since it has a treatment effect.
- $\mathbf{C}_8$: Race of the subjects. Coded as a binary variable, white and Other.
- $\mathbf{C}_9$: Tumor size, continuous variable.

In order to built the Cox regression model the proportional hazards assumption was tested for the number of positive nodes, which is the exposure of interest. The null hypothesis was

- H_0 : The hazard is proportional for the range of the exposure

The proportionality test, yielded a p-value of 0.851, meaning that we cannot reject the null hypothesis and the hazards are proportional for the exposure.

3.2 Meta-analysis on IPD

In the Cox model, when using the IPD, modelling the exposure with restricted cubic splines, we chose three knots and estimated two coefficients describing the risk on the outcome. The position of the knots were at 0,2 and 15 positive nodes for the range of the exposure. These points were chosen based on the distribution of the positive nodes, that featured about 90% of the subjects having up to 5 positive nodes and only 1046 with more than 15 positive nodes. After acquiring the two coefficients for each study,

describing the dose-response relation for the exposure of interest, the studies were allowed to follow their own shape and study-specific predictions were made for a selected range of the exposure, 0 to 50. The studies were then pooled at each level. To allow for heterogeneity the random effect model was considered. In figure 3.1 the dose-response shape for the 9 studies can be seen along with the overall shape acquired from the point-wise meta-analysis of the studies.

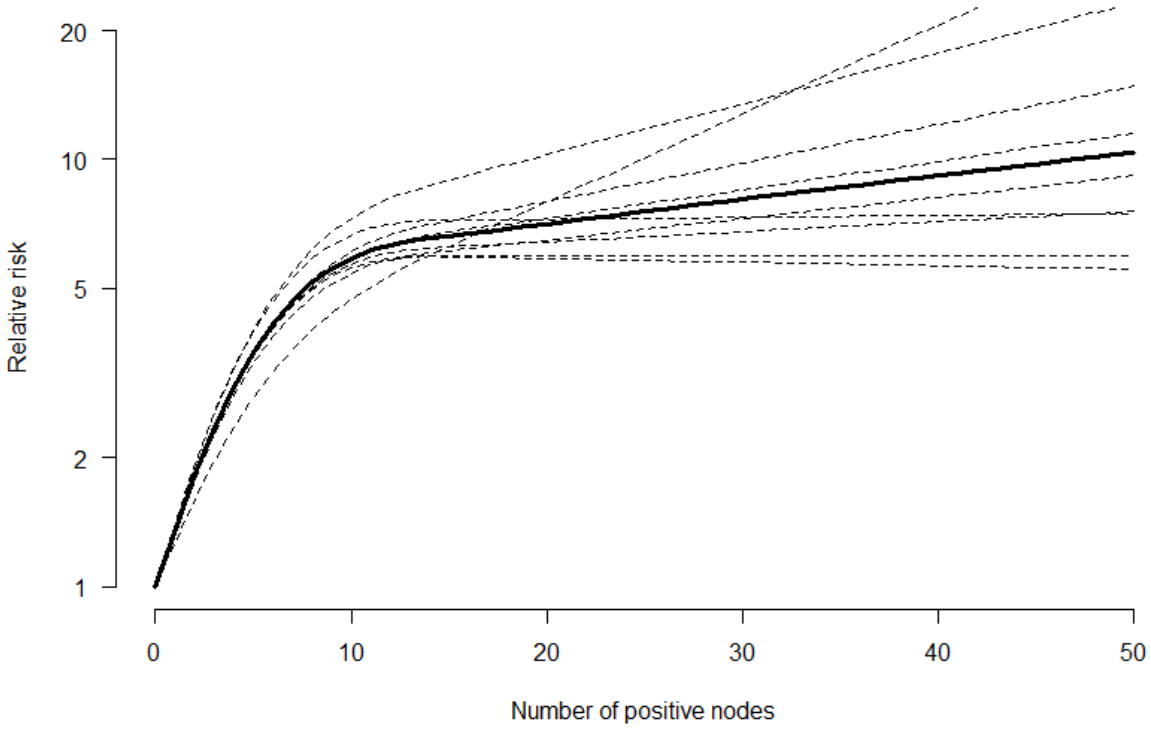

Figure 3.1: Dose-Response relation of the IPD in the nine registries (dashed lines) overlaid with the overall curve from the point-wise meta-analysis, modelled with restricted cubic splines

The same analysis for the dose-response relation was also replicated using second order fractional polynomials to model the exposure, estimating two coefficients describing the risk on the outcome. The Cox regression model

was used, stratifying by studies, and the overall best choice of p_1, p_2 was selected. All the possible 36 combinations of p_1, p_2, were considered and the model with the lowest AIC value was chosen. In figure 3.2 the study-specific shapes can be seen along with the overall shape acquired from the point-wise meta-analysis of the studies, considering a random effects model.

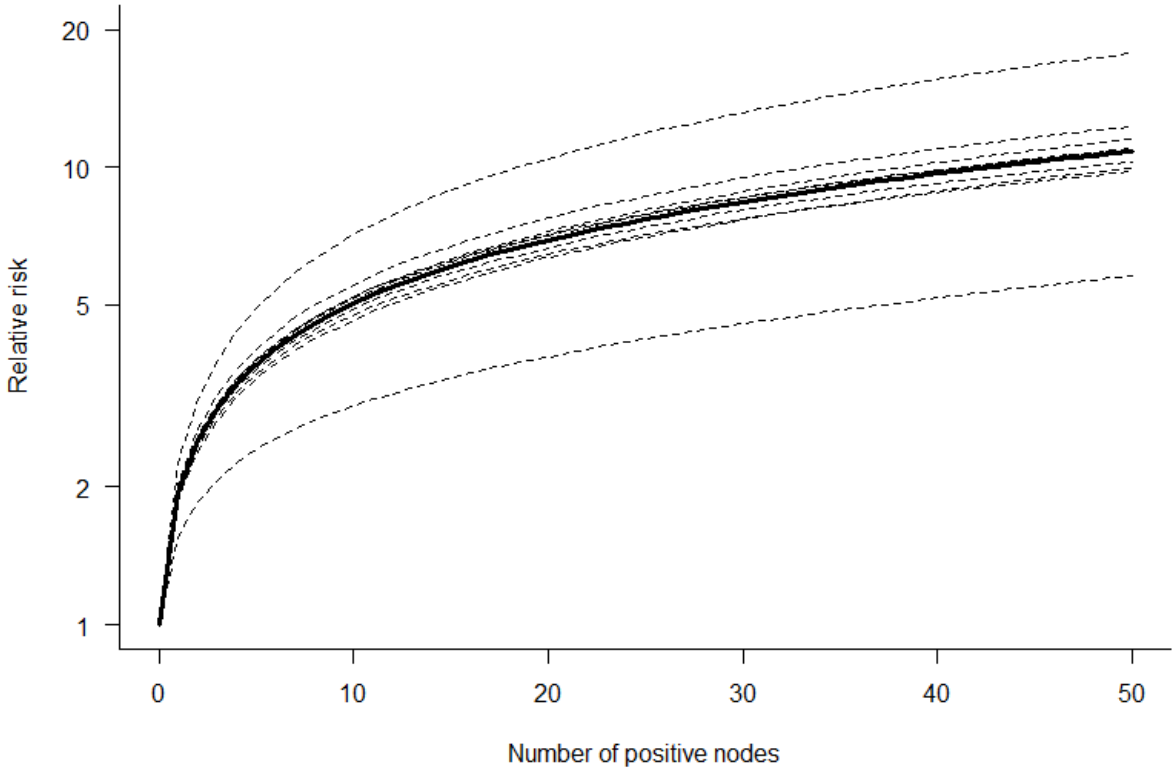

Figure 3.2: Dose-Response relation of the IPD in the nine registries (dashed lines) overlaid with the overall curve from the point-wise meta-analysis, modelled with second order fractional polynomials

3.3 Summarized data

In addition to the analysis performed for the IPD, the dataset was also summarized for the 9 studies. The exposure of interest was divided into the following categories based on its distribution: 0-1 positive nodes, 2-4 positive nodes, 5-24 positive nodes and 25-50 positive nodes. Using the same Cox model as in the last setting, but in this case without a transformation for the exposure, the risks for the 4 categories were obtained, relative to the first category, along with the median of each category and the respective standard errors and confidence intervals for the estimates of the relative risks. Using these, we created the usual tabular form as we would expect from a published study. A snapshot of the aggregated data for one study is seen in table 3.2.

Table 3.2: Cumulative Incidence data of the San Francisco-Oakland study on positive nodes and breast cancer mortality rate

Positive nodes	Dose	Cases	Total Patients	Person Years	Rate	log(HR)
0-1	0.0	713	11419	75734.67	0.007	0.00
2-4	2.0	276	1624	9960.42	0.047	0.82
5-24	8.0	393	1198	6517.25	0.060	1.45
25-50	28.0	17	29	129.17	0.131	2.23

3.3.1 Meta-analysis using restricted cubic splines

In the analysis of the aggregated dataset, we used the log-linear model from equation 2.1. The position of the knots was chosen to be the same for both the IPD and the aggregated data for comparison reasons, although, that the same justification clearly applies for the choice of the knots. For the aggregated dataset the point-wise approach was implemented as described

in section 2 of the thesis, considering a random effects model. In figure 3.3 the study-specific dose-response shapes can be observed for all 9 studies for a selected range of the exposure overlaid with the overall shape from the point-wise meta analysis.

The results of both the analyses, comparing the IPD with the summarized data, can be seen in figure 3.4. A direct comparison of the two shapes shows that the results drawn from summarized data, as far as the dose-response relation is concerned, very similar to the results of the same analysis on IPD. The estimated relative risks for selected points of the exposure is shown in table 3.3 along with the confidence intervals, both for the analysis in the IPD and the analysis on the summarized data.

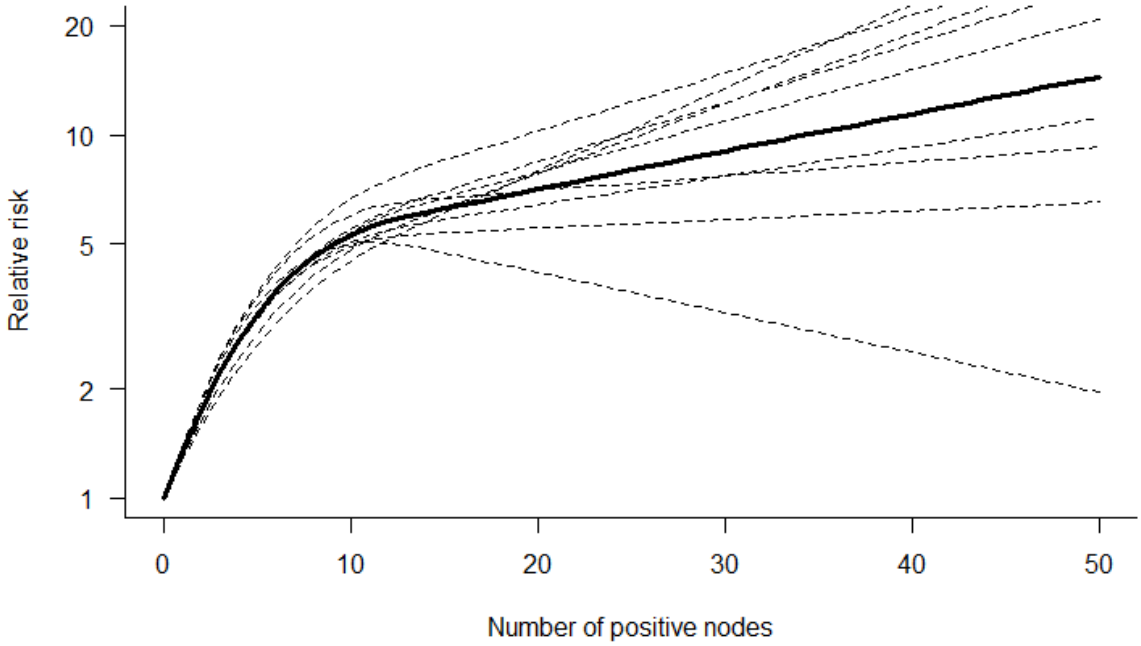

Figure 3.3: Study specific curves for the nine studies included in the meta-analysis. The number of positive nodes was modelled using restricted cubic splines (dashed lines). The thick line represents the pooled curve of the point-wise meta-analysis.

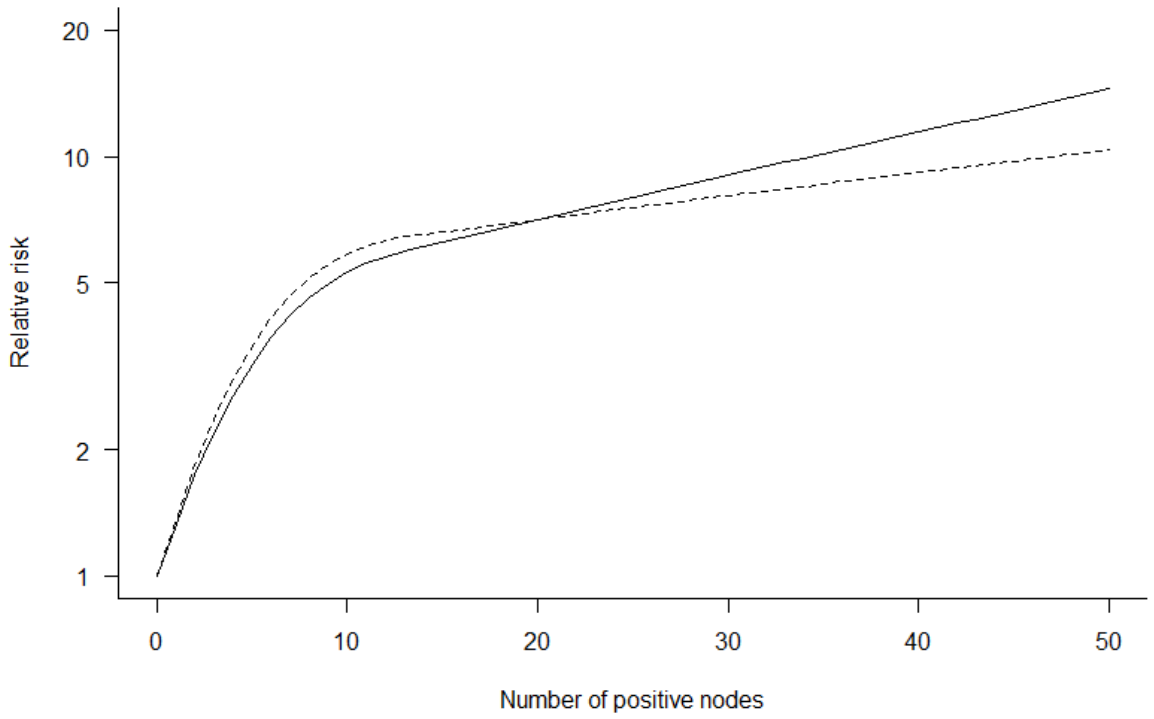

Figure 3.4: Comparison of the point-wise meta-analysis on the IPD with the point-wise meta analysis on the summarized data using restricted cubic splines to model the exposure. The straight line represents the point-wise meta-analysis on summarized data and the dashed line the point-wise meta-analysis on the IPD.

Table 3.3: Predicted relative risks (RR) of the point-wise approach of IPD and the point-wise approach on summarized data

Dose	RR from Summarized data	95% CI	RR from IPD	95% CI
0	1.00	(1.00,1.00)	1.00	(1.00,1.00)
1	1.33	(1.31,1.35)	1.37	(1.35,1.39)
2	1.74	(1.69,1.78)	1.83	(1.78,1.88)
5	3.20	(3.06,3.35)	3.53	(3.36,3.71)
10	5.30	(5.00,5.61)	5.87	(5.53,6.23)
15	6.26	(5.71,6.87)	6.62	(6.18,7.08)
20	7.09	(6.10,8.23)	7.06	(6.39,7.81)
30	9.02	(6.87,11.85)	8.07	(6.70,9.71)
40	11.43	(7.66,17.07)	9.16	(6.97,12.02)
50	14.47	(8.52,24.60)	10.36	(7.23,14.85)

3.3.2 Meta-analysis using fractional polynomials

A second order fractional polynomial was also used to model the exposure for the aggregated data applying it to the model from equation 2.1. The strategy used to choose the best model for each study was the choice of the model that minimizes the AIC value (Rota et al. [4]). In the analysis of the summarized data, all possible 36 combinations for each study were considered again, only that this time the p_1, p_2 minimising the AIC value was selected within each study, allowing for different choice of p_1, p_2 for the different studies. We can see the optimal choice of p_1, p_2 for each study in table 3.4 together with the corresponding AIC value.

Table 3.4: Overall choice of the within studies combination of p_1, p_2 that minimizes the AIC value

Study	p_1	p_2	**AIC value**
San Francisco-Oakland	0.5	3	-1.1368
Connecticut	0	1	-3.9950
Metropolitan Detroit	-0.5	1	-3.5051
Hawaii	-0.5	-0.5	-4.2395
Iowa	-1	1	-0.1893
New Mexico	0	3	-3.4718
Seattle (Puget Sound)	-0.5	0	-0.0244
Utah	-2	0	-3.3812
Metropolitan Atlanta	-2	-2	-1.5904

The dose-response relation shapes for the 9 registries can be seen in figure 3.5, together with the overall shape on the meta-analysis, considering a random effects model. The results of the meta-analysis on the aggregated data set, compared to the analysis on the IPD can be seen in figure 3.6 where we notice that both methods produce similar dose-response relations, especially in low number of positive nodes, where about 90 % of the subjects lay. In table 3.5 we can see the estimated relative risks for selected points of the exposure. We note that both the restricted cubic splines and the fractional polynomial regression methods to model the exposure generate accurate results, compared to the same analysis of the IPD.

Furthermore, both methods yield very similar results, although the curves in figure 3.6 using fractional polynomials not as "steep" as when modelling when restricted cubic splines seen in figure 3.4. In figures 3.6 and 3.4 the two lines deviate for high levels of the exposure, something that can be explained due to the low number of subjects and events at these levels.

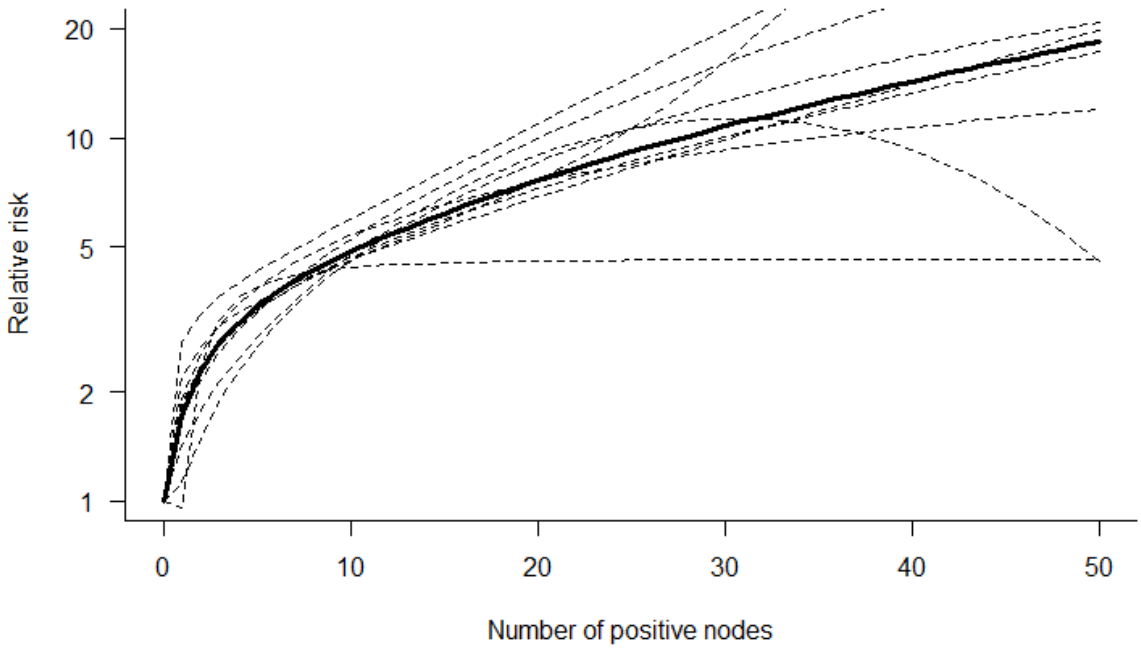

Figure 3.5: Study specific curves for the nine studies included in the meta-analysis. The number of positive nodes was modelled using fractional polynomials (dashed lines). Study-specific p_1, p_2 were chosen according to the minimum AIC values. The thick line represents the pooled curve of the point-wise meta-analysis.

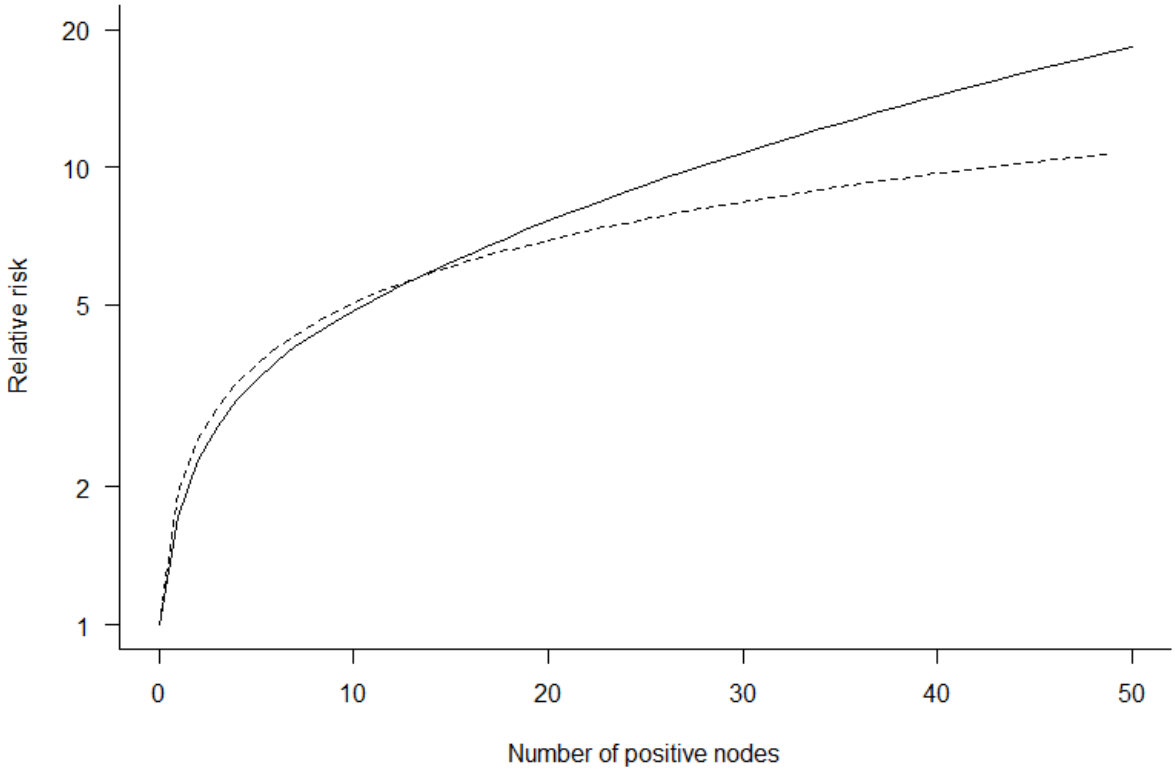

Figure 3.6: Comparison of the meta-analysis on the summarized data (straight line) with the point-wise meta analysis on the IPD (dashed line) using fractional polynomials to model the exposure

Table 3.5: Predicted relative risks (RR) of the point-wise approach of IPD and the point-wise approach on summarized data

Dose	RR from Summarized data	95% CI	RR from IPD	95% CI
0	1.00	(1.00,1.00)	1.00	(1.00,1.00)
1	1.72	(1.44,2.053)	1.95	(1.88,2.03)
2	2.28	(2.01,2.59)	2.53	(2.39,2.67)
5	3.42	(3.17,3.69)	3.71	(3.44,3.99)
10	4.86	(4.59,5.14)	5.04	(4.61,5.51)
15	6.19	(5.53,6.94)	6.07	(5.51,6.68)
20	7.63	(6.39,9.12)	6.94	(6.27,7.68)
30	10.77	(7.82,14.82)	8.42	(7.56,9.37)
40	14.33	(8.87,23.14)	9.69	(8.68,10.81)
50	18.39	(9.79,34.52)	10.73	(9.61,11.99)

3.4 Heterogeneity at different dose-levels

Based on the point-wise approach discussed in section 2.6, the vector I^2 for heterogeneity at each point of the exposure was also calculated for the summarized dataset when modelling the exposure with restricted cubic splines. The plotted results can be seen in figure 3.7. The values before 12-13 positive nodes might seem a bit odd, something that is explained by the fact that I^2 is forced to take positive values. As a result, for low values where more that 95 % of the subjects lay, there is minimal heterogeneity and the I^2 value is set to 0. This can be also verified by figure 3.3, where we notice that all the studies present similar dose-response association for low levels of the exposure. Although for higher levels the studies present somewhat different results, the heterogeneity reduces due to the low precision of the study-specific predictions at those levels.

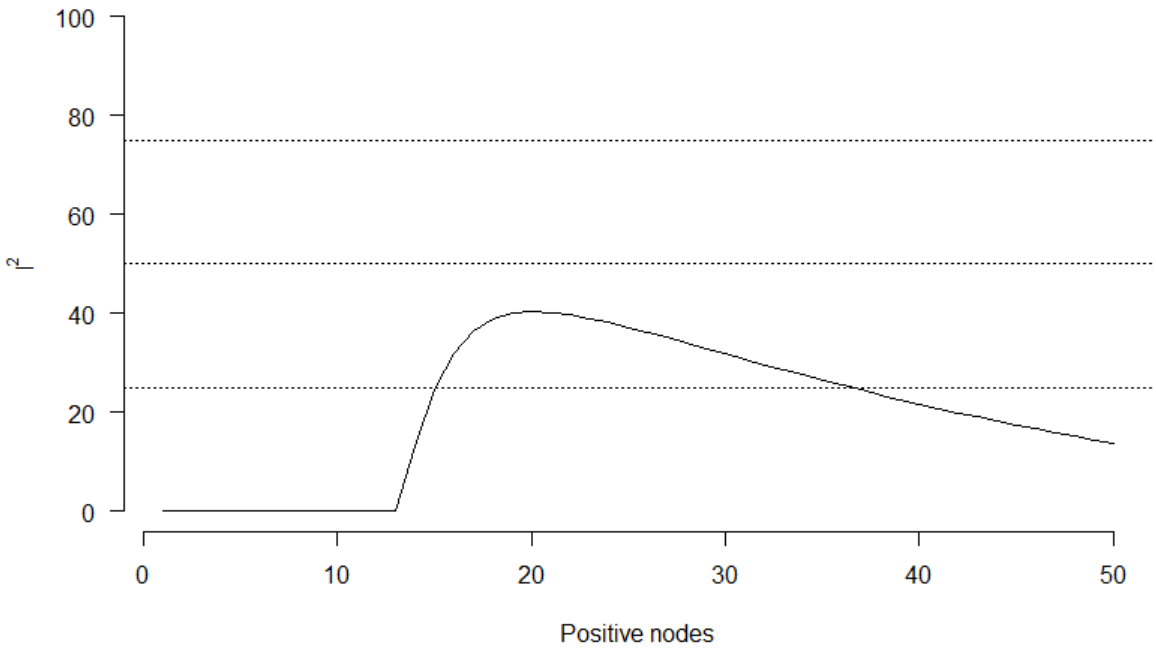

Figure 3.7: Point-wise heterogeneity of the studies estimated when modelling with restricted cubic splines

3.5 Simulated studies

Following the results from the analysis on the individual patient data, we evaluated the performance of the point-wise approach in different dose-response shapes, other than the one described in figures 3.6 and 3.4 ("plateau" shape). The shapes considered are the most common in dose-response meta-analysis , specifically: U-shape relation, J-shape relation and linear relation. In order to perform the analysis in these different scenarios, we simulated data for the different dose-response relations for the three different shapes as follows:

1. Define the dose-response relation (linear,J-shape, U-shape) of the population. The range of the exposure desired to be investigated was defined, between 0 and 12. Then three exposure level points were chosen, $\mathbf{x} = (x_1, x_2, x_3)$, within the range. A vector of probabilities $\mathbf{p} = (p_1, p_2, p_3)$ of the outcome was chosen corresponding to the values of x, based on the equation of a logistic regression model. Analogous to the vector $\mathbf{p}$, in connection with the logistic regression model a $\mathbf{y} = (y_1, y_2, y_3)$ vector of log(odd) was created, where each element y_i is calculated as

$$y_i = log\left(\frac{p_i}{1 - p_i}\right) \tag{3.2}$$

Finally having calculated the log(odds) corresponding to the probabilities of the outcome the following equation was solved to estimate the coefficients that would describe the desired dose-response relation, given they rise from a model with a quadratic form

$$\mathbf{Y} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \tag{3.3}$$

so that with Y and X known the vector of coefficients is calculated as

$$\boldsymbol{\beta} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{pmatrix}^{-1} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \tag{3.4}$$

Solving the equation 3.4 the results would be a vector of three coefficients, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ defining the coefficients of a logistic regression model, where the first coefficient represents the intercept of the model and the last two the quadratic dose-response relation. Following the calculations as depicted above, the probability of the outcome given the dose is calculated

$$p = \frac{e^{log(odds)}}{1 + e^{log(odds)}} \tag{3.5}$$

where the log(odds) at a specific value of the dose are defined as

$$log(odds) = \beta_1 + \beta_2 dose + \beta_3 dose^2 \tag{3.6}$$

2. Draw exposure distributions from a lognormal distribution for 10 different studies. Each study was allowed to have a different number of observations varying from 3.000 to 10.000. The number of subjects for each study was randomly chosen.

3. Based on the exposure value, simulate the outcome based on the probabilities calculated from equation 3.5. After extracting the simulated outcome, we derive two vectors for each simulated study, one with the level of the dose and one with the outcomes, sufficient information to define the dose-response relation.

4. Categorize the exposure distribution randomly, into either 3 or 4 intervals for each study.

5. Estimate a logistic regression model on the aggregated data and create a summarized dose-response dataset.

6. Based on the procedure described in section 2, perform the pointwise meta-analysis. The exposure was modeled using restricted cubic splines with three knots placed at the $10th$, the $50th$ and the $90th$ quantiles of the distribution.

As an illustration, the dose-response U-shape relationship is presented in figure 3.8. Since we notice that the lowest risk to experience the event is at the dose level of 6, we use it as the reference point in the dose-response analysis.

Meta-analysis was also performed on the IPD, running a logistic regression model on each study to extract the study-specific trend coefficients, which

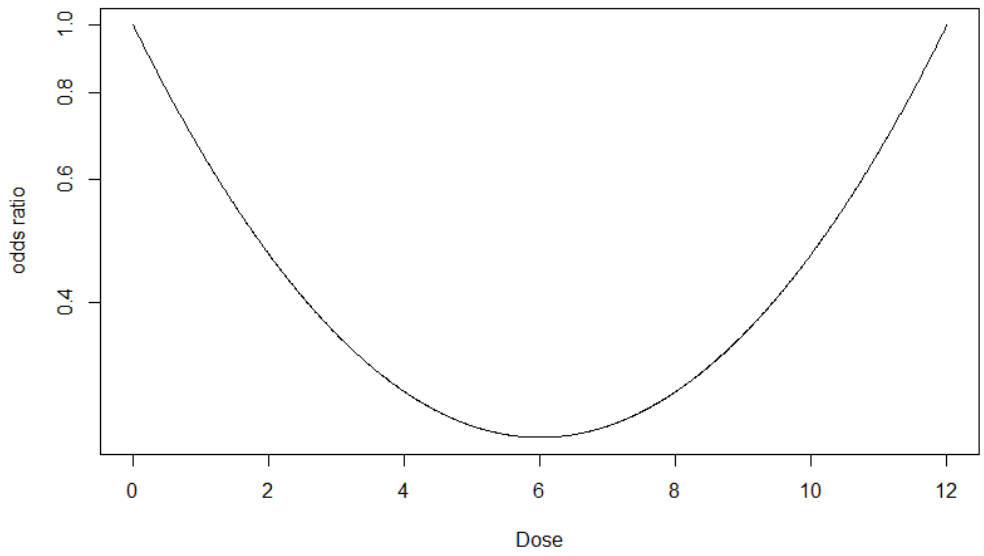

Figure 3.8: Simulated U-shape dose-response relation

were then used for study-specific predictions. The data used were the vectors of doses and events as defined in step 3. The studies were pooled at the levels of the exposure that was also modeled using restricted cubic splines with the same three knots as in the dose-response meta-analysis of summarized data. It is our interest to explore whether the performance of the point-wise approach on the summarised data would be similar to the performance of the meta-analysis of IPD, which is considered to be the golden standard.
In figures 3.9, 3.10 and 3.11 the results for the analyses can be seen. We can notice there that both the analysis with IPD and with summarized data (dashed line and straight line respectively) give similar results and compared to the true risk (dotted line) the estimations of both methods are very accurate.

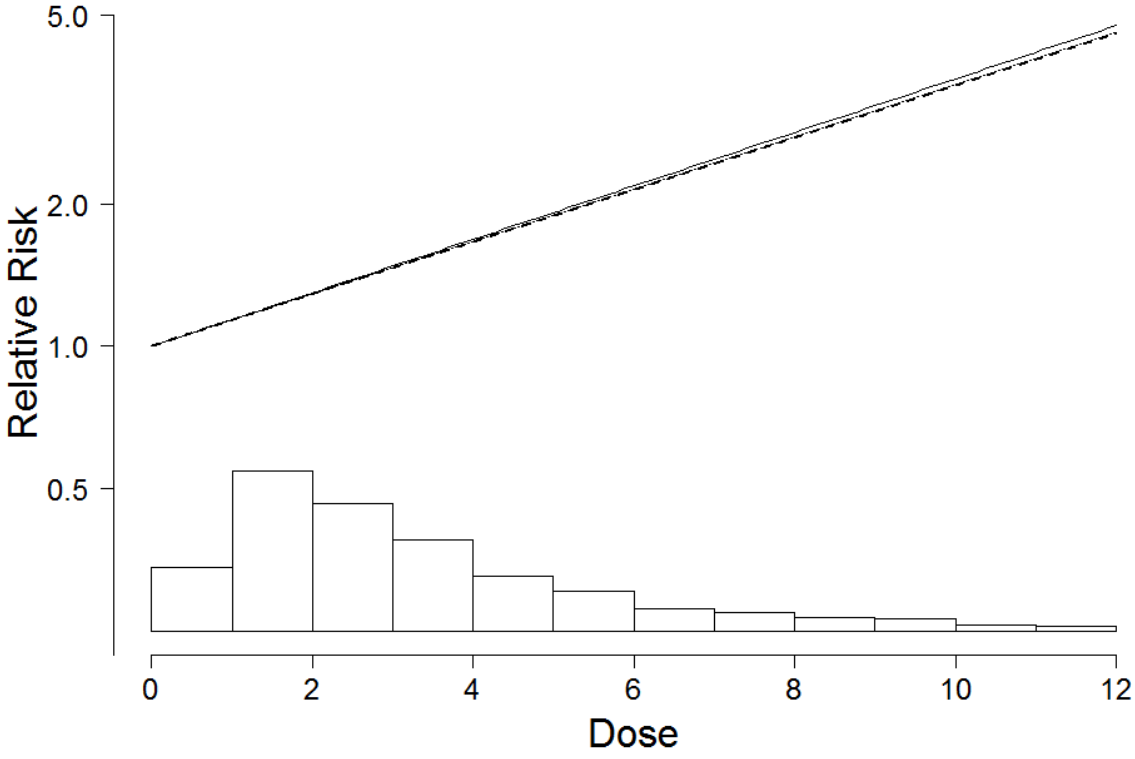

Figure 3.9: Comparison based on simulated studies plotted together with the distribution of the data. The point-wise approach on summarized data (solid line) is compared to the point-wise approach on IPD (thick dashed line) and the true underlying shape (dotted line) for a linear trend analysis

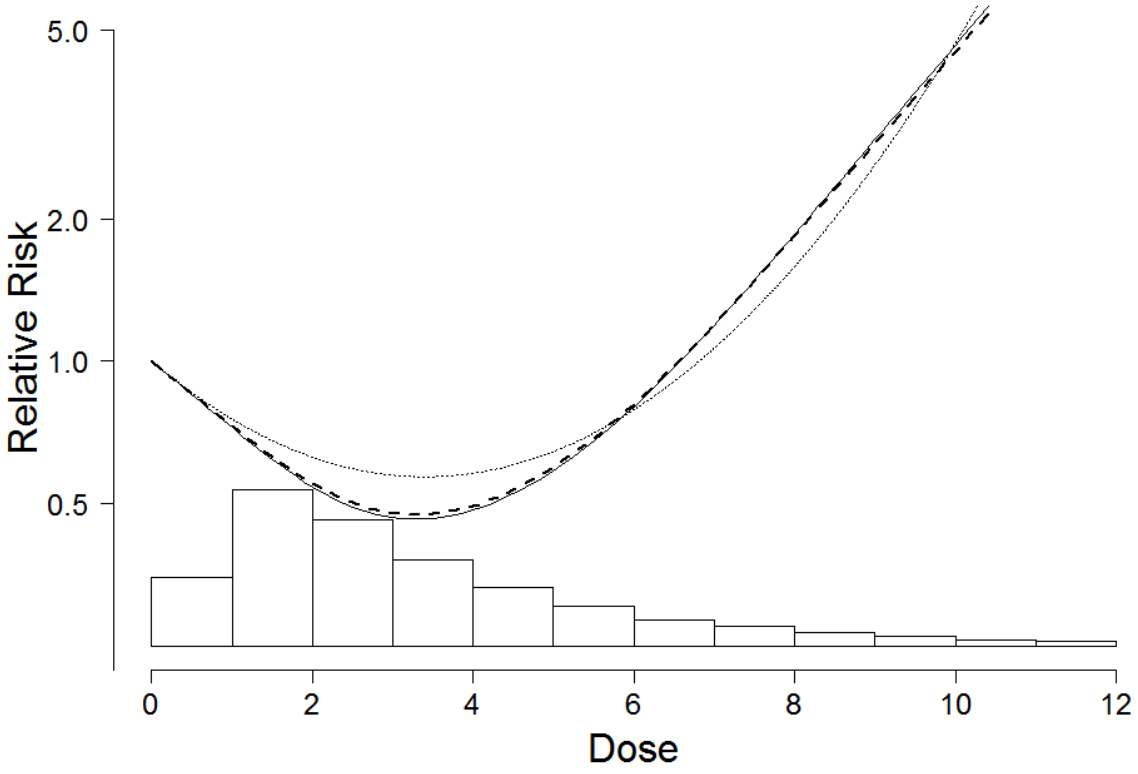

Figure 3.10: Comparison based on simulated studies plotted together with the distribution of the data. The point-wise approach on summarized data (solid line) is compared to the point-wise approach on IPD (thick dashed line) and the true underlying shape (dotted line) for a J-shape relation

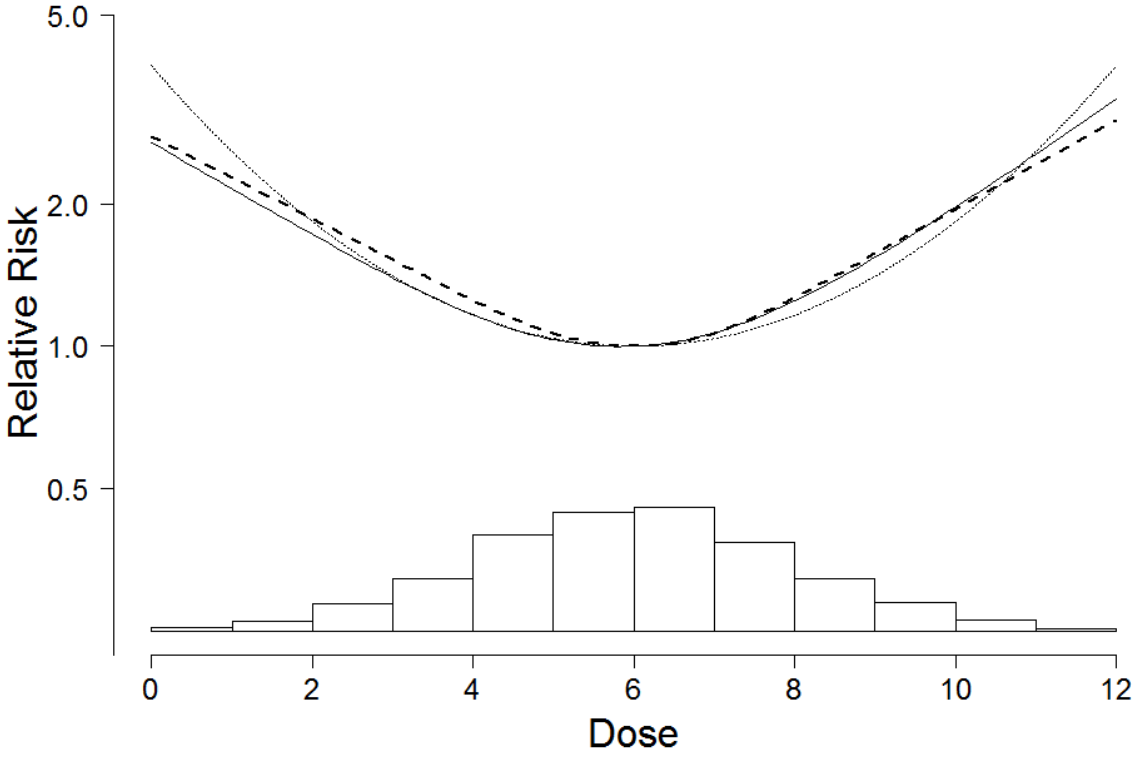

Figure 3.11: Comparison based on simulated studies plotted together with the distribution of the data. The point-wise approach on summarized data (solid line) is compared to the point-wise approach on IPD (thick dashed line) and the true underlying shape (dotted line) for a U-shape relation

Chapter 4

Discussion

In this thesis we discuss and demonstrate the application of the point-wise approach in the context of dose-response meta-analysis of summarized data.

This approach comprises of three main steps, which have been thoroughly demonstrated in the analyses performed. Firstly, the trend coefficients of each study are estimated, using either restricted cubic splines or second order fractional polynomials to model the exposure, to allow non-linear shape of the dose-response curve. Then, based on the coefficients extracted, study specific predictions are made for a desired range of the exposure, estimating the relative risks of the outcome relative on a suitable common reference point. Finally, for each level of the exposure, the estimations of the relative risk are averaged, with different weight according to a fixed or random effect model.

The methodology is described in detail in chapter 2 and is applied to survival data from the SEER registry involving breast cancer patients. A further application was demonstrated on simulation studies with common dose-response shapes tested. The ability of the method to fit dose-response relations was evaluated in comparison with the meta-analysis on IPD. In all the analyses the results of the novel method were similar to the results obtained with the analysis on IPD, which is considered the golden standard in meta-analysis. Such results, demonstrate that the novel approach can be a useful tool, when access to the IPD is not possible.

When modelling with restricted cubic splines, the positions of the knots is chosen based on the distribution of the exposure. In our case, since the distribution of the exposure in the SEER dataset was highly right skewed the choice of the knots was in low values. In the simulates simulated studies the $10th$, the $50th$ and the $90th$ quantiles of the exposure distribution were chosen as the position of the knows. When modelling with second order fractional polynomials, the strategy for the best choice of the power vector of the transformations, is the choice of the power vector that minimizes the AIC value of the model.

Limitations commonly encountered in meta-analysis of summarized data that have not been addressed in this thesis are publication bias, different patterns of selecting subjects in each study, the ranges of the exposure examined, or confounding in observational studies.

In our analyses with the IPD the exposure was categorised in the same way in all studies. This is not usual in practice, as studies can use different categories. As heterogeneous exposure can be a complication, in the simulated studies the exposure was randomly categorised. The results of the analysis using the point-wise approach demonstrate that the number of categories and the doses assigned does not effect the performance of the proposed approach.

The analyses implemented were carried out in the statistical software R, available for free on CRAN. An updated version of the *dosresmeta* package will be made available to include the point-wise methodology as an alternative option for dose-response meta-analysis of aggregated data.

Based on our findings, we conclude that the proposed methodology can give very accurate results and it can be trusted to perform dose-response meta-analysis on summarized data. This is particularly useful when access to the original data is not available.

In summary, we believe that the method proposed in this thesis will improve the overall quality and practice in reporting the findings of quantitative reviews of summarized dose-response data. Although the results obtained from the developed method look promising, further analyses need

to be made and more cases need to be studied in order to improve the precision of the method and further develop its accuracy.

Bibliography

[1] Sander Greenland and Matthew P Longnecker. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *American Journal of Epidemiology*, 135(11):1301–1309, 1992.

[2] Vincenzo Bagnardi, Antonella Zambon, Piero Quatto, and Giovanni Corrao. Flexible meta-regression functions for modeling aggregate dose-response data, with an application to alcohol and mortality. *American Journal of Epidemiology*, 159(11):1077–1086, 2004.

[3] Nicola Orsini, Ruifeng Li, Alicja Wolk, Polyna Khudyakov, and Donna Spiegelman. Meta-analysis for linear and nonlinear dose-response relations: examples, an evaluation of approximations, and software. *American Journal of Epidemiology*, 175(1):66–73, 2012.

[4] Matteo Rota, Rino Bellocco, Lorenza Scotti, Irene Tramacere, Mazda Jenab, Giovanni Corrao, Carlo La Vecchia, Paolo Boffetta, and Vincenzo Bagnardi. Random-effects meta-regression models for studying nonlinear dose–response relationship, with an application to alcohol and esophageal squamous cell carcinoma. *Statistics in Medicine*, 29(26):2679–2687, 2010.

[5] Willi Sauerbrei and Patrick Royston. A new strategy for meta-analysis of continuous covariates in observational studies. *Statistics in Medicine*, 30(28):3341–3360, 2011.

[6] Jian Qing Shi and JB Copas. Meta-analysis for trend estimation. *Statistics in Medicine*, 23(1):3–19, 2004.

[7] Qin Liu, Nancy R Cook, Anna Bergström, and Chung-Cheng Hsieh. A two-stage hierarchical regression model for meta-analysis of epidemiologic nonlinear dose–response data. *Computational Statistics & Data Analysis*, 53(12):4157–4167, 2009.

[8] A Gasparrini, B Armstrong, and MG Kenward. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Statistics in Medicine*, 31(29):3821–3839, 2012.

[9] Ilias Thomas, Alessio Crippa, and Nicola Orsini. A point-wise average approach in dose-response meta-analysis of summarized data for binary outcomes. *Manuscript*, 2015.

[10] Elizabeth L Turner, Joanna E Dobson, and Stuart J Pocock. Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. *Epidemiologic Perspectives & Innovations*, 7(1):9, 2010.

[11] Alessio Crippa and Nicola Orsini. Multivariate dose-response meta-analysis: the dosresmeta r package. *Manuscript*, 2015.

[12] Nicola Orsini, Rino Bellocco, Sander Greenland, et al. Generalized least squares for trend estimation of summarized dose-response data. *Stata Journal*, 6(1):40, 2006.

[13] Jesse A Berlin, Matthew P Longnecker, and Sander Greenland. Meta-analysis of epidemiologic dose-response data. *Epidemiology*, 4(3):218–228, 1993.

[14] Michael Borenstein, Larry V Hedges, Julian Higgins, and Hannah R Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2):97–111, 2010.

[15] Julian Higgins and Simon G Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558, 2002.

[16] Sylvain Durrleman and Richard Simon. Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561, 1989.

[17] Loic Desquilbet and François Mariotti. Dose-response analyses using restricted cubic spline functions in public health research. *Statistics in Medicine*, 29(9):1037–1057, 2010.

[18] Frank E Harrell. *Regression modeling strategies*. Springer Science & Business Media, 2001.

[19] Vincent Vinh-Hung, Tomasz Burzykowski, Jan Van de Steene, Guy Storme, and Guy Soete. Post-surgery radiation in early breast cancer: survival analysis of registry data. *Radiotherapy and Oncology*, 64(3):281–290, 2002.

[20] Patricia Tai, Gábor Cserni, Jan Van De Steene, Georges Vlastos, Mia Voordeckers, Melanie Royce, Sang-Joon Lee, Vincent Vinh-Hung, and Guy Storme. Modeling the effect of age in t1-2 breast cancer using the seer database. *BMC Cancer*, 5(1):130, 2005.

Appendix

R code

Example code to perform point-wise dose-response meta-analysis on summarised data when modelling with restricted cubic splines. The example is based on the SEER dataset to get figure 3.3.

```
## read the dataset
tabDataSum <- read.table("http://www.imm.ki.se/biostatistics/data/seer.txt",
                         header=TRUE)

## define the knots
knots <- c(0, 2, 15)

## Running spline model within each study
modiSum <- lapply(split(tabDataSum, tabDataSum$id), function(x)
   dosresmeta(formula = logrr ~ rcs(dose, knots), type = "ir", cases = cases,
              se = se, n = n, id = id,  data = x)
)

## define the range of the exposure
newdata <- data.frame(dose = seq(0, 50, 1))

## Study specific predictions (and SE) for the range of the exposure
predSum <- lapply(modiSum, function(x)
   data.frame(predict(x, newdata, expo = F, se = T)[, c(1, 3, 6)])
```

```
)

## Dose-specific predictions (and SE) for the studies
## data to be used for meta-analysis
predDosesSum <- lapply(as.list(seq(nrow(newdata))), function(x)
   data.frame(t(sapply(predSum, function(y) y[x, ])))
)[-1]


## Dose-specific meta-analysis (as many as the points in newdata)
metamodiSum <- lapply(predDosesSum, function(x)
   rma.uni(y = unlist(x$pred), sei = unlist(x$se), method = "REML")
)


## Final results: Dose-specific predictions (and SE)
newpredSum <- data.frame(do.call("rbind", lapply(metamodiSum, function(x)
   predict(x, transf = exp))))


## Plot of final results
par(mfrow = c(1, 1))
with(newpredSum,{
   plot(newdata$dose, c(1, pred), log = "y",
        type = "l", las = 1, bty = "l", col = "black", ylim = c(1, 20),
        ylab = "Relative risk", las = 1,
        xlab = "Number of positive nodes", lwd=3)
})


## Add study-specific shapes
lapply(modiSum, function(x)
```

```
  with(predict(x, newdata, expo = T),{
    lines(get("rcs(dose, knots)dose"), pred, lty=2, type="l"
    )
  })
)
```