INDIVIDUAL PATIENT DATA META ANALYSIS OF PROGNOSTIC FACTOR STUDIES

By

GHADA MOHAMMED ABDALLAH ABO-ZAID

A thesis submitted to
The University of Birmingham for the Degree of
Doctor of Philosophy

School of Mathematics
The University of Birmingham
September, 2011

Abstract

Prognostic factors (PFs) are patient characteristics (e.g. age, biomarker levels) that are associated with future clinical outcomes in patients with a disease or health condition. Evidence-based PF results are paramount, for which individual patient data (IPD) meta-analysis is thought to be the 'gold-standard' approach, as it synthesises the raw data across related studies (in contrast to an aggregate data meta-analysis, that just uses reported summary data).

In this Ph.D. thesis, I investigate statistical issues and develop methodological recommendations for individual patient data meta-analysis of prognostic factor studies (IMPF) projects. First, I investigate the benefits and limitations of IPD meta-analyses of PF studies through a systematic review and in-depth evaluation of existing IPD meta-analyses of PFs; 48 IMPF articles were found and an in-depth evaluation of a random sample of 20 IMPF articles was undertaken to identify how such projects are initiated, conducted, and reported, and to identify the benefits and challenges of the IPD approach. I found that although IMPF articles have many advantages, they still face a number of challenges and pitfalls such as different methods of measurements, ignoring clustering of patients across studies, missing data, and potential publication bias, unachieved linearity assumption of PFs, poor reporting, and potentially not protocol driven. To improve IMPF articles and projects guidelines were developed, and an array of methodological research questions identified.

Secondly, I undertook an empirical study to compare between the IPD and aggregated data approach to assess PFs in breast cancer. I showed that the IPD approach is preferable over aggregated data, as it allows one to adjust the PF by other confounding factors, examine PFs in subgroups of patients and assess the interaction between two PFs as an additional PF. It also allowed more studies and more patients to be included. However, the IPD approach still faced challenges, such as potential publication bias, missing data, and failed model assumptions in some studies.

Thirdly, I developed eleven IPD meta-analysis models to investigate whether accounting for clustering of patients within studies should be undertaken and which approach is the best to use. The models differed by using either a one-step or two-step approach, and whether they accounted for parameter correlation and residual variation. An IPD meta-analysis of 4 studies for age as a PF for 6 month

mortality in traumatic brain injury was used as an applied example. Surprisingly, I found that there was *no* difference between the eleven models because there was little variation in baseline risk across studies. Thus, a simulation study was undertaken to examine which model is the best one-step or two-step, and whether accounting for the clustering of patients within studies is important. I found that the clustering across studies should be considered, and one-step model accounting for the clustering of patients within studies is the best fitted model as it yielded the lowest bias and the coverage was around 95%. Ignoring clustering can produce downward bias and too low a coverage; occasionally the two-step produces too high a coverage.

Fourthly, I examined the linearity assumption for the relation between age and risk of 6 months mortality in the traumatic brain injury dataset. I found that the linear trend was not the best in all studies. Thus, I developed three *non-linear* fractional polynomial IPD meta-analysis models based on whether one-step or two-step approach and whether first or second order fractional polynomial functions are performed. I found that one-step fractional polynomial meta-analysis model that account for the clustering of patients within studies is again the best fitted model, as it easier to fit and force the IPD studies to have the same polynomial powers. This revealed age has a quadratic relationship with mortality risk.

Fifth, I assessed whether small-study effects (i.e. potential publication bias) exists for 6 IPD prognostic factor articles by using different tools, such as contour funnel plot, cumulative meta-analysis, trim and fill method, and regression tests. I found the small-study effects problem is not a major concern, in contrast to aggregated data meta-analysis of PFs. Only in the breast cancer data of Look et al. was there substantial evidence for small-study effects. However, adjusted results to account for this provided a smaller PF effect but suggested the original conclusions are unlikely to change.

To sum up, this thesis highlights a number of challenges of IMPF projects and discusses possible approaches to dealing with some of them. However, numerous challenges remain for future work.

Acknowledgments

Firstly, I am heartily thankful to my supervisor, Richard Riley, whose encouragement, supervision and support from the preliminary to the concluding level enabled me to develop an understanding of the subject. I am particularly grateful for how he has helped develop this career path for me, and for the large amount of time he has spent reading various drafts and providing constructive feedback over the last few months, without his guidance and persistent help this dissertation would not have been possible. I also want to greatly thank Prof. Jon Deeks for their encouragement, guide, and advice throughout the duration of the thesis, especially for the final stage in my thesis, it was really helpful. Furthermore, I thank Prof. Willi Sauerbrei who provided helpful direction and feedback at some important stages of the thesis. I also would like to thank Dr. Prakash Patil, Boliang Guo and Sergi for their help and advice in some particular area in my thesis.

I would like to show my gratitude to my sponsor, Embassy of the Arab Republic of Egypt cultural centre and Educational Bureau London, for awarded me with this scholarship.

I would like also to thank the MRC Midlands Hub for Trials Methodology Research for their support and provision, especially Prof. Lucinda Billingham for her support, help and advice that she gave to me.

I owe my deepest gratitude to my parents, my brothers and my sister for their support and encouragement throughout the duration of the thesis. Last but not the least, I offer my regards and blessings to all of those who supported me in any respect during the completion of the thesis.

LIST OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

PF	Prognostic Factor
IPD	Individual Patient Data
IMPF	Individual Patient data Meta-Analysis of Prognostic Factor
uPA	Urokinase type Plasminogen activator
OR	Odds Ratio
TBI	Traumatic Brain Injury
IMPACT	International Mission for Prognosis and Analysis of Clinical Trials in Traumatic Brain Injury
OS	Overall Survival
RFS	Relapse Free Survival
HR	Hazard Ratio
Ln(HR)	Log Hazard Ratio
Ln(OR)	Log Odds Ratio
C.I.	Confidence Interval
FP	Fractional Polynomial
FP1	First Order Fractional Polynomial
FP2	Second Order Fractional Polynomial

CHAPTER 1

AIMS AND BACKGROUND FOR THE THESIS

1.1 Introduction

In healthcare, prognosis is the study of future health related events in patients with a particular disease or condition. Prognosis research studies can take a prospective or retrospective approach. In prospective studies, patients are recruited at the current date and followed for an adequate length of time to identify what outcomes are achieved. This allows baseline characteristics and outcomes to be recorded by the researcher. On the other hand, retrospective studies utilise existing data (e.g. from cancer databases) about the baseline characteristics and outcome. Most prognosis studies are retrospective due to the ease and speed of using existing data. However they are restricted by the data collected (e.g. certain variables and outcomes may be missing)[1]. Prospective studies are preferred as they allow greater conduct on the data collected and the quality of the study (e.g. complete follow-up, less missing data).

There are many different types and objectives of prognosis studies, but most usually have the aim to identify which factors measured at some baseline point (e.g. diagnosis of disease), are associated with the outcome, or predict the outcome. However, other types of prognosis studies might look to predict response to treatment, or ascertain the overall prognosis and survival rates in certain diseases. Such distinct objectives of prognosis research are not well understood[2], and section 1.2.2 unpacks this further.

Prognostic factors (PFs) are patient characteristics (e.g. age, biomarker levels) that are

associated with future clinical outcomes in patients with existing disease. PFs are important clinical tools because they help to identify patients with different risks of outcome (e.g. recurrence of disease) and thereby facilitate the most appropriate treatment strategies and aid patient counselling[3], they are also used in the design and analysis of trials, and act as confounding factors in observational studies[4]. Evidence regarding PFs is therefore very important to both clinicians and their patients. However, primary PF studies have numerous problems such as poor reporting, analysis and design[5-7]. PF studies are subject to numerous biases, such as selective reporting[7] and "optimal" choice of cut-points[8]. These severely limit meta-analysis of PF studies using aggregated data, where summary of results are combined across PF studies. Individual Patient Data Meta-analysis of PF studies (IMPF) have been proposed as the gold-standard approach[9-11], because it utilises the raw patient data and thus does not rely on reported results.

The aims of this thesis are to examine the feasibility of an IMPF articles and to tackle a number of statistical and methodological challenges when undertaking IMPF projects. I begin with a broad discussion on the different types of prognosis research and then introduce the fundamental concepts for primary PF studies and their analysis, and the rationale and basic methodology for meta-analysis of PF studies.

1.2 Types of prognosis research

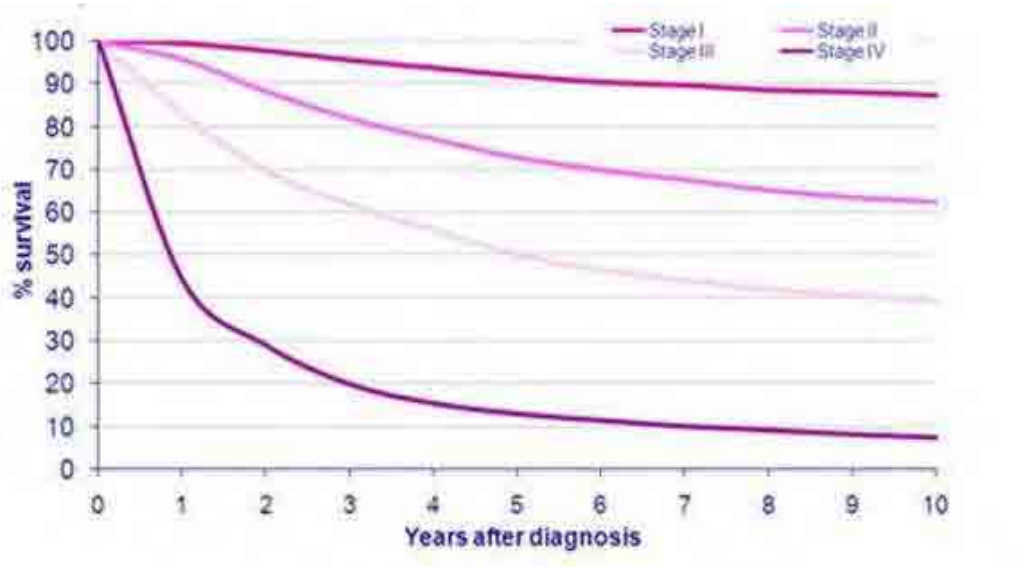
There are four types of prognosis research:

1.2.1 Overall prognosis

An often forgotten part of prognosis research is to establish the natural prognosis in a disease. For example, it is of interest to know survival rates at various time points following diagnosis of

breast cancer. This may be sub-grouped further by characteristics (such as pre to post menopausal), prognostic factors (see below), and certain treatments. In some poor countries, it may unfortunately be possible to establish natural prognosis of patients with a disease, where no treatment is available.

Figure 1.1: 0-10 year relative survival for breast cancer by stage, diagnosed in the West Midlands 1990-1994 followed up to the end of 2004, as at December 2008[12]



The overall prognosis of a condition is important for looking at trends in, say, survival rates over time (Figure 1.1), planning and preparing for future healthcare demands in the population (e.g. if patients live longer with disease), and understanding 'baseline risk' so to help see if any new treatment and PFs modify it. For example, Figure 1.1 shows 0-10-year relative survival for breast cancer by stage, diagnosed in the West Midlands 1990-1994 followed up to the end of 2004[12].

1.2.2 Prognostic Factors

Prognostic factors (PFs) are patient characteristics or measurements (e.g. age, biomarker levels)

that are associated with future clinical outcomes in patients with existing disease[13]. For example, Figure 1.1 shows that stage of disease is a PF, as the survival curve is different for different stages. PF studies investigate the prognostic value of one or more particular variables; for example, in breast cancer Look et al.[14] Investigate whether Urokinase type Plasminogen Activator (uPA) and its inhibitor PAI-1 were associated with overall and disease free survival. PF studies are the most common type of prognosis study. Hundreds of papers are published each year in oncology alone[15].

PFs have many uses; for example, they are the building block for prognostic models (see below), and are potential confounding factors in observational studies and randomised trials to be adjusted for. They can also be incorporated in the design and analysis of randomised trials, to ensure treatment groups are balanced (via stratified sampling) and to increase power (by including them as covariates in the analysis).

As a detailed example of a PF study consider Braun et al.[16], whose primary objective was to detect whether there is an association between the presence of micro-metastasis at diagnosis in the bone marrow with survival or recurrence of disease in breast cancer patients; i.e. whether micro-metastasis is a PF in breast cancer at diagnosis. The authors adjusted for other PFs (e.g. age, menopausal status, tumour size, tumour grade) to see whether the presence of micro-metastasis in the bone marrow was still a significant independent PF after accounting for these other factors. They also assessed micro-metastasis as a PF in sub-groups of patients according to those who received endocrine treatment alone or chemotherapy alone; they found that it is still a PF. In this thesis the focus is on using meta-analysis to identify PFs by combining across multiple PF studies like Braun et al.[16]

1.2.3 Prognostic Models

Prognostic models, also called prediction models, risk scores or prognostic scores, evaluate many variables simultaneously to determine a model that predicts the outcome for the individual patient, based on the final set of included variables. A prognostic model is crucially different to a PF; PFs may be associated with the outcome, but the predictive ability of each PF for the individual patient may be poor. A well-known example for prognostic model is the Nottingham Prognostic Index (NPI), which is used to predict mortality risk in an individual patient with breast cancer based on a set of factors such as tumour size and tumour grade[17].

Figure 1.2 and 1.3 shows why a PF is usually not an accurate predictor of absolute risk for the individual patient[18]. In particular, Figure 1.2 shows Kaplan-Meier survival curves for each of three groups of patients defined by “low”, “medium” and “high” levels of a PF; the distinct curves show that this factor is prognostic as it is associated with differential risks of outcome.

Figure 1.2: Kaplan-Meier survival curves for each of three groups of patients defined by 'low', 'medium', and 'high' levels of a prognostic factor[18]

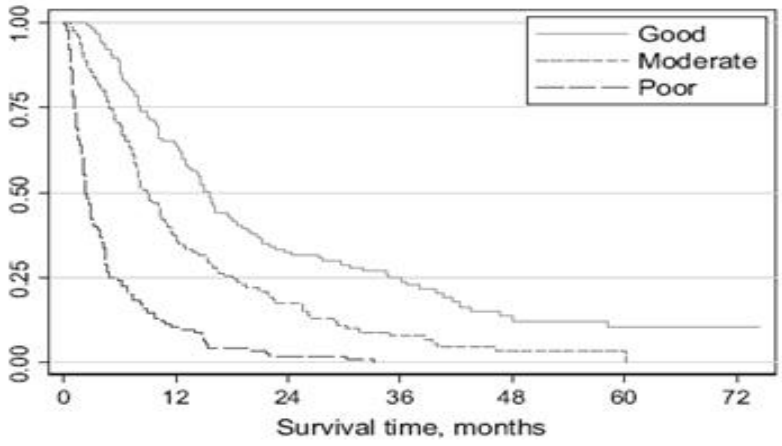
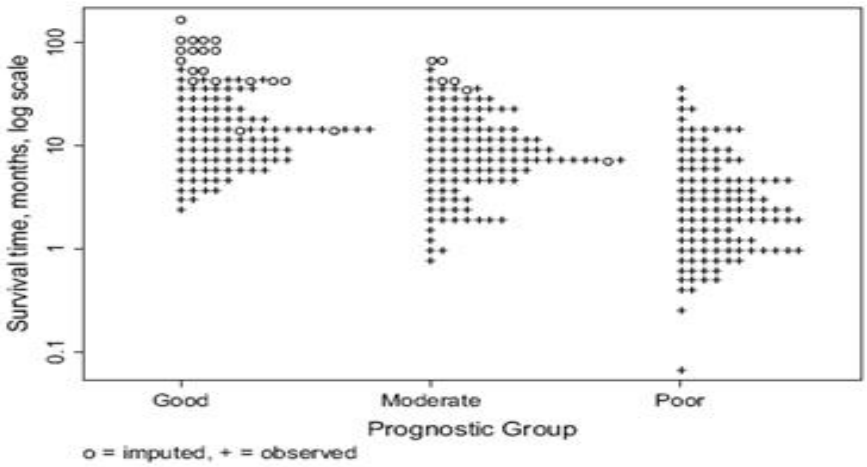


Figure 1.3 shows the distribution of survival times for the same patients in the same study amongst individual patients in each group defined by three levels of a PF. There is wide overlap of

the distributions, emphasizing that levels of the PF do not accurately predict survival times for the individual patient.

Figure 1.3: The distribution of survival times, from the same patients in the same study is shown amongst individual patients in each group defined by three levels of a prognostic factor[18]



Thus, prognostic models are concerned with individual predictive accuracy, which is usually improved by using multiple PFs together. This difference between PFs and prognostic models is often not understood, and in particular researchers often interpret a PF as if it can be used for making individual predictions, when this has actually not been evaluated. Note that prognostic models are not the focus in this thesis.

1.2.4 Predictive Factors

Predictive factor studies are concerned about identifying variables that predict treatment response; in cancer these are known as predictive markers, though this terminology is not wide-spread outside the cancer field. For example, breast cancer studies often investigate what variables modify response to adjuvant hormone therapy or polychemotherapy[2]. Predictive factor

studies essentially look to detect treatment covariate interactions or treatment effects in subgroups, often within a randomized trial design. For instance, oestrogen receptor status in breast cancer is a predictor of responsiveness to treatment of hormonal therapy, but it contributes little information on prognosis otherwise (i.e. it is not a PF). While progesterone receptor status is a PF but it does not influence treatment response[2]. So clearly not all predictive factors are PFs. Predictive factors will not be discussed further in this thesis.

1.3 Prognosis Research - A Turbulent Time

Although there is continuously a large amount of research activity, very few prognosis studies seem to impact upon clinical practice because PF studies still have problems such as missing data, different methods of measurements, publication bias and availability bias. Further, increasing evidence suggests that prognosis research is often poorly designed, poorly reported, and subject to numerous biases such as selective reporting and publication bias. Hemingway et al.[2] outline many challenges required to improve quality and impact of prognosis research which are now summarized below.

Purpose of the studies

Often the goals of prognosis research studies are not clear. In particular researchers often perform a PF study and conclude it can be used to decide treatment strategies or predict likely outcomes for the individual patients, when this has not actually been investigated. Thus the awareness of the main goals of prognosis research and the distinction between them needs to be better understood by prognosis researchers. There is a need to firmly establish taxonomy of the goals and types of prognosis research. The PROGRESS initiative (a UK initiative to improve Prognosis research) has been formed to tackle this problem, and currently distinguishes the

four types of study as outlined in section 1.2.

Funding

There is a lack of strategic framework for funding prognosis research[2]; as a result many prognosis studies have a "*what's in the freezer*?" approach where existing data, often of poor quality, is given to a junior (who may have no training in prognosis research) to data dredge for prognostic results. This causes many problems, not least publication bias and selective outcome reporting. Thus, there is a challenge here to guide funders to set up a strategic framework for investment toward different types of high-quality prognosis research. This undoubtedly requires the goals and impact of prognosis research to be better exemplified, so that funders can clearly see the benefits.

Protocol

Protocols are a very important first step in any healthcare research project. Yet, in many current prognosis research studies there is no protocol[5]. At the study onset the protocol should mention the research aims, inclusion, and exclusion criteria of patients, outcomes assessed and the statistical analysis plan. Yet, this is often not done, highlighting again the "*what's in the freezer approach*[2] ". Thus, encouraging protocols for prognosis research should be seen as a high priority.

Methods

Current evidence highlights poor quality in the methodology of prognosis research[2] [19]. Methodology can be improved by developing and using appropriate methods in study design, statistical analysis, and reporting[20]. In the study design the research should begin with

a clear and well-defined research question so that their study can be designed and targeted appropriately[19]; in particular prospective protocol driven studies are desired with large sample size and adequate power. In Holländer et al.[6] the authors showed that the choice of the statistical methods has strong influence on the results and on the interpretation of PF studies. So it is important to use the most reliable methods. This thesis will consider statistical methods at the meta-analysis level.

Reportingly, there is currently a lack of sufficient information in PF study reporting[15 21]. For example, many important statistics are not reported[15 22] (e.g. the number of patients in each group, the number of events in each group, hazard ratios, confidence intervals, P-values) , which makes it hard to reliably interpret PF studies and include them in a meta-analysis (see later).

For improvement, developing generic reporting guidelines is important. A good start toward this goal are the REMARK guidelines[23]. These are aimed at prognostic tumour marker studies in cancer, but are generally applicable to all PF studies in other disease, see Altman et al[24] and Riley et al.[22], for other reporting guidelines for prognosis and survival studies.

A particular problem is that researchers often look at a wide range of PFs and outcomes, but selectively only report significant outcomes and factors[8]. Thus the researchers should have to pre-specified PFs and outcomes for the project before starting the study; this again relates to the need for a protocol.

1.4 Prognostic Factor Research

Section 1.2 outlined four key areas of prognostic research. From now on this thesis focuses on the research and analysis of PFs.

PF studies are perhaps the most common prognosis study, and they are desperately in need of improved research standards. Current evidence shows that PF studies are generally poorly designed, use small sample sizes, are poorly analyzed and reported, and are subject to numerous biases including publication bias, selective reporting and biased choice of cut-off level[22 23 25] .

As briefly mentioned in section 1.2, there are many reasons why PFs are important. They can help to: (i) identify the causal pathway (e.g. what factors are causally associated with poor outcome); if such factor are modifiable (e.g. smoking, obesity), then clinicians can help prevent poor outcome; (ii) identify variables to consider within prognostic models; (iii) identify factors that may confound results in observational study and (unbalanced) randomized trials; (iv) identify factors to consider in the design of randomized clinical trials (e.g. standardized randomization) and analysis (e.g. to increase power by adjusting for PFs); (v) inform what subgroups to look at within overall prognosis studies (e.g. age is a PF in traumatic brain injury, so producing survival by age groups can aid patient and family counselling).

1.4.1 Types of Prognostic Factor Studies

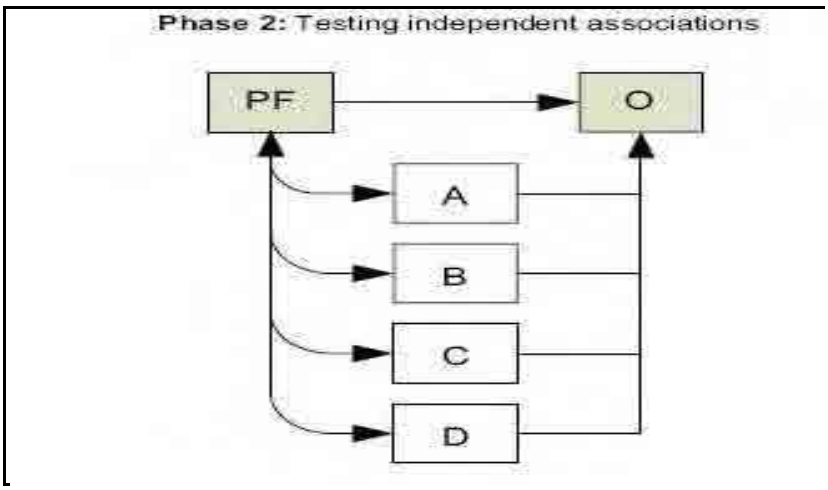
Hayden et al.[26] suggest PF studies can also be called 'explanatory studies' as they look to find factors associated with explaining outcomes. Hayden outlines three phases of PF studies (Phase I, Phase II, and Phase III).

In Phase I, studies identify the association between a number of potential factors and outcomes. For example, in lung cancer, a Phase I study might assess smoking status as a PF alongside many other PFs such as age, gender, biomarker level...etc. Thus Phase I studies do not

study one specific PF but look at many PFs of interest to explore which appear associated with outcome.

In Phase II, studies test the independent PF value of a small number of PFs identified in Phase I research. In other words, these studies aim to measure the independent effect of a particular PF (or a small number of PFs) while controlling for other PFs. For example in Look et al.[14] the authors aimed to measure the independent prognostic effect for uPA and PAI1 after adjusted by existing PFs (e.g. hormone receptor, lymph nodes, age) in relation to overall survival, see Figure 1.4.

Figure 1.4: Phase II- testing the independent association of PF adjusted by other variables



Hayden (2007) et al.[26] suggest Phase III as studies are "*those that attempt to describe the complexity of the prognostic pathways or processes. These studies apply knowledge from the previous phases of study on independent associations and incorporate other knowledge from the field of study*". These studies include the PF of interest, existing PFs, and other variables that are thought to be intermediate in the pathway toward the outcome, and the outcome of interest. Thus, Phase III studies essentially look at the role of PFs on the causal pathway to disease

outcome. Note that the focus in this thesis is on Phase I and Phase II PF studies. These are the most common published PF papers in the medical literature, and look to establish whether one or more factors are prognostic, but are not concerned with understanding the causal pathway or whether a PF is indeed causal. Most PFs, such as age, are not actually causal themselves, but associated with other (unknown) factors or underlying causal pathways.

1.4.2 Types of Data in Prognostic Factor Studies

PF data are either continuous or binary. For example, in breast cancer markers uPA and PAI-1 are continuous variables and hormone receptor is a binary variable. However continuous PF variables are often dichotomized into a 'low' and 'high' group (e.g. uPA is often dichotomized by choosing a cut-off point)[8 27]. This is not recommended because it does not maintain the whole information of the variable, reduces power and leads to inaccurate results[28 29]. Note that other types of data are possible for PFs and outcomes (e.g. ordinal), but the focus in this thesis is only on continuous or binary PF data, and time until the event occur and binary data for outcome(s).

Binary Outcome Data

Outcomes in PF studies are often binary, with two possible values (one and zero) representing whether the outcome occurred or not (e.g. dead =1, alive = 0). For example, in Hukkelhoven et al.[30] the authors assess the association between age (per 10 years) and 6 month mortality (dead = 1, alive = 0) in patients with severe traumatic brain injury (TBI) using data from 4 trials. Table 1.1 details the IPD available within Hukkelhoven et al.[30] with each patient in each study providing their age (defined as age in years divided by 10) and whether they were dead or alive at six months. This database has been kindly provided by E. Steyerberg[31] for use

in this thesis.

(N.B. a set of IPD from multiple studies often contains thousands of patients. This is the case in Table 1.1, so for brevity I do not show all the rows of data here. Rather I use four horizontal dashes to indicate where non-displayed rows of data occur, and these follow in a similar manner to the displayed rows above them).

Table 1.1: IPD from 4 trials that assess whether age is a PF for six month mortality on TBI, from Hukkelhoven et al.[30]

Study id	Patient	Age (years/10)	Event(dead = 1, alive = 0)
1	1	1.9	0
1	2	2.5	0
1	3	2.7	0
-	-	-	-
-	-	-	-
-	-	-	-
1	825	3	1
2	1	1.6	0
2	2	2.7	0
2	3	2.1	0
-	-	-	-
-	-	-	-
-	-	-	-
2	959	5.3	0
3	1	2.5	1
3	2	8.9	1
3	3	6.1	1
-	-	-	-
-	-	-	-
-	-	-	-
3	466	2.2	1
4	1	2.3	1
4	2	2.2	1
4	3	5.9	0
-	-	-	-
-	-	-	-
-	-	-	-
4	409	4	1

Survival Data

Another common type of outcome data in PF studies is time to event or survival data. Survival data is the time it takes for a certain event to occur (time to event), (e.g. time to death or time to recover from a certain disease). In the breast cancer study of Look et al.[14] the authors assessed whether uPA and its inhibitor (PAI-1) are PFs. The study included 18 IPD datasets provided by member of the European Organization for Research and Treatment of Cancer Receptor and Bio-marker Group (EORTC-RBG), including 8837 breast cancer patients. The association of many potential PFs (including Lymph nodes status, hormone receptor status, menopausal status, age, uPA, PAI1) with overall survival (OS) and relapse free survival (RFS) were assessed. The IPD from 15 studies have been kindly provided for use in this thesis by Look et al.[14] Data are listed in Table 1.2 for some selected PFs for the 15 studies.

PF studies are often time to event studies, as they follow patients from a baseline point (where PFs are measured) over time until an event occurs or the study ends or the patients leave the study, which I illustrate in Figure 1.5.

Figure 1.5: Time until the events occurs

Table 1.2: Survival data for IPD from 15 studies, assessing PFs in breast cancer, Look et al[14]. 2002

Study No.	Study Name	Patient id	OS	Event(dead = 1, alive = 0)	uPA	PAI-1	Hormone Receptor	Lymph nodes status
1	Rdam	1	13.73	0	0.81	53.85	1	1
1	Rdam	2	120.00	0	0.59	14.46	1	1
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
1	Rdam	2722	69.78	0	0	5.43	1	1
2	Utr	1	78.69	0	0.02	0	1	0
2	Utr	2	93.80	0	2.93	6.23	0	0
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
2	Utr	199	78.85	0	0.10	0	1	1
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
15	Swiss	1	31.97	0	0.65	6.90	1	0
15	Swiss	2	39.56	0	0.35	2.10	0	0
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
15	Swiss	663	55.33	0	0.58	4.50	1	1

Note that: OS is Overall Survival; uPA is Urokinase type Plasminogen activator and its inhibitor PAI-1

The length of time from baseline to the event is called the survival time. Patients who do not experience the event have a censored survival time. The type of censoring shown here is called right censoring, where the true event occurs after (to the right of) the censoring time. Other forms of censoring can occur, such as left censoring[32], but in this thesis the focus is on the right censoring data, as this is the most common for PF studies.

Figure 1.6 Example of censored and uncensored data for 6 patients followed over time

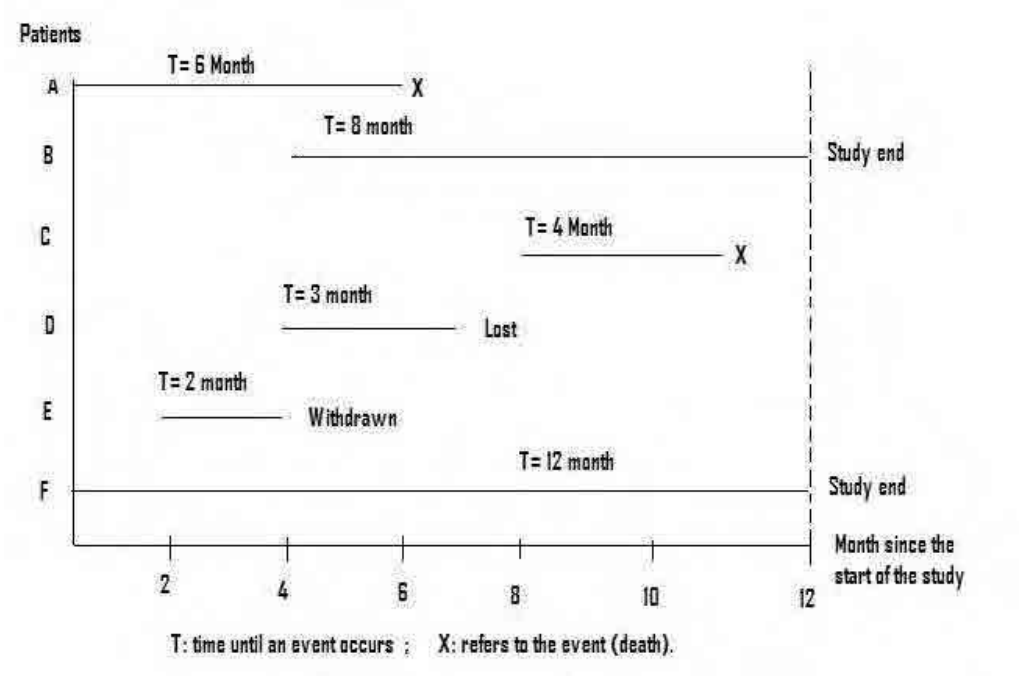


Figure 1.6 illustrates the experience for 6 patients followed over time. The horizontal axis represents time since the start of the study in months and the vertical axis represents patients. X refers to the event (death). Both Patient A and Patient C are uncensored observations, as the time of occurrence the event (death) is known. On the other hand, Patients B, D, E, and F are censored, as the time of the occurrence of the event (death) is unknown. For example, Patient B was recruited at month 4 but by the end of the study the event had not occurred. Patient D started

at month 4 but he was lost to follow up after 3 months. Usually, the time of entry into the study is not important, and so all patients are set to a $T = 0$ baseline point and their survival times give the time from zero to their event or censoring.

1.5 Logistic Regression Analysis in a Single Study

Researchers of PF studies must analyse data using the most appropriate statistical methods. The choice of statistical model depends firstly on the type of outcome, which for PF studies is usually binary (e.g. dead / alive) or time to event (time to death or censoring) as discussed, and this thesis focuses on these two outcome types. In the next section, I briefly consider basic statistical models for binary data and time to event data, to set the foundation for subsequent chapters. Note I consider here a single PF study; meta-analysis of multiple PF studies is considered in section 1.7.2.

1.5.1 Odds Ratio

In a simple situation of a binary outcome (e.g. dead / alive), and a binary PF (e.g. Hormone receptor), a two by two table can be constructed to look at the PF effect on outcome.

Table 1. 3: Two by two table for prognostic factor and outcome

	Outcome	
PF	Dead	Alive
Positive hormone receptor	A	b
Negative hormone receptor	C	d

For example, one can calculate an estimate of the odds ratio (OR) and its confidence interval As follows:

$$\text{Odds ratio} = \frac{ad}{bc} \tag{1.1}$$

This gives the odds of death in the positive group versus the odds of death in the negative group, with OR value greater than one indicating a higher odds for positive patients.

The 95% confidence interval is

$$Ln(OR) \pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \tag{1.2}$$

However, to assess a continuous PF or to allow multiple PFs to be examined together a more advanced method such as logistic regression is needed.

1.5.2 Logistic Regression

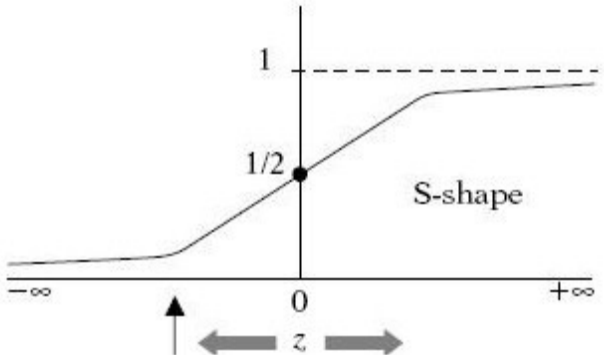
Logistic regression is a frequently used generalized linear model[33], which can be used to describe the relationship between a set of variables and the probability of a binary dependent variable. Specifically, it models the probability transformed by the logistic function as follows:

$$\boldsymbol{P(z)} = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e^{-z}}} \tag{1.3}$$

Where $P\,(z)$ is between zero and one; it is equal to zero when $z = -\infty$, and equal to one when $z = \infty$. The logistic function allows an S-shape such that the influence of z on the probability P is minimal for low values until some threshold is reached. Then, probability P increase over a certain range of intermediate z values, until it increases rapidly toward one

once z gets large enough. Then, P remains extremely close to one, see Figure 1.7. The model is useful for assessing PFs because z can be replaced by a linear function of (unknown) parameters relating to the PFs under consideration as follows:

Figure 1.7: S-shape for various values of z



$$P = \frac{1}{1 + e^{\propto + \sum_{i=1}^{k} \beta_i X_i}} \tag{1.4}$$

Or equivalently,

$$Ln\left(\frac{p}{1-p}\right) = \propto + \sum_{i=1}^{k} \beta_i X_i \tag{1.5}$$

Where X_i refers to the set of $i = 1\ to\ k$ PFs under considerations; α and β_i are unknown parameters to be estimated. α denotes the *ln-odds* of outcome probability for patients with all PFs equal to zero, where odds is defined by$\left(\frac{p}{1-p}\right)$. β_i gives the *ln-odds* ratio between two patients who differ in patients $X\acute{}i$ by one unit, with all other X_i values $(i \neq i\acute{})$ the same; the odds ratio is thus $exp(\beta_i)$. So, for a binary PF such as hormone receptor (positive = 1, negative = 0) in breast cancer, the odds ratio gives the odds of the outcome for the positive hormone receptor relative to the odds for patients with negative hormone receptor after adjusting for other PFs in the model. For a continuous PF such as age, the odds ratio gives the ratio of odds for two patients who differ in age by a value of 1 unit.

1.5.3 Application of Logistic Regression

As an illustrated example, I now apply logistic regression to one study (study 1), within IPD provided by Hukkelhoven et al.[30] in order to assess the association between age and 6 month mortality in patients with severe traumatic brain injury [Table 1.1]. Six month mortality is coded as 1 if a patient died and 0 if the patient was still alive at six months. Age is the one PF of interest here, denoted by X and it is a continuous variable. As the effect of a one year increase in age is likely to be small, in accordance with the original authors, I coded age in years dividing by 10, so to model the effect of a 10 year increase in age. The study includes 825 patients and 199 died. The logistic regression model can be written as follows:

$$Ln\left(\frac{p_k}{1-p_k}\right) = \propto + \beta\, X_k \qquad (1.6)$$

Where p is the probability of death at 6 months for patient k. The model was fitted by using maximum likelihood method for estimating the unknown parameters, within the STATA package, and the results are shown in Table 1.4.

Table 1. 4: The estimation result for the age and traumatic brain injury after 6 month for binary data

Parameters	Parameter	Standard error	Confidence Interval		P-value
			Lower	Upper	
α	-2.20	0.24	-2.67	-1.73	<0.001
β	0.31	0.06	0.19	0.45	<0.001
$Odds\ ratio = \exp(\beta)$	1.36	--	1.20	1.55	<0.001

Table 1.4 shows that increasing age significantly increases the probability of mortality by 6 months. The odds ratio for two patients who differ in age by 10 years is estimated as $exp(0.31) = 1.36$; which indicates that the odds of mortality increase by 36%, for a 10 years increase in age, with a 95% confidence interval of 1.20 to 1.55. Thus we conclude from this analysis that age is associated with 6 month mortality, and thus age is a PF in traumatic brain injury patients. This is a simple illustrate analysis. Extensions to nonlinear trends and adjustment for other PFs will be considered in later chapters, within a meta-analysis setting.

1.6 Survival Analysis of Prognostic Factor in a Single Study

1.6.1 The Survival Function

The analysis of time to event data assesses the survival function, denoted as S(t), where

$$S(t) = P(T > t) \tag{1.7}$$

With P meaning probability and T is the random variable for a patient's survival time. T can be any number greater than or equal to zero. $S(t)$ is the probability that the patient survives longer than a specified time t. For example, if $t = 2$ years, then the survival function will be $S(2) = P(T > 2)$ which means the probability a patient survives longer than two years after the baseline point (usually diagnosis of disease). In relation to this thesis, it is of interest to identify PFs that modify $S(t)$.

1.6.2 The Hazard Function

The hazard function is denoted as $\lambda(t)$ which can be defined as the instantaneous probability at time t for having the event at that time. It can be written mathematically as follows:

$$\lambda(t) = \lim_{\delta \to 0} \frac{P(t \leq T \leq (t + \delta t) | T \geq t)}{\delta t} \tag{1.8}$$

In other words, hazard function is defined as the conditional probability that the patient's event time T will lay in the time interval between t and $t + \delta t$, given that the event does not occur before time t. The hazard function is nonnegative, taking any value from 0 to ∞ and can start from any value and increase or decrease over time.

The relationship between hazard function and survival function is given as follows:

$$S(t) = exp\left[-\int_0^t \lambda(x)dx\right] \tag{1.9}$$

$$\lambda(t) = -\left[\frac{S'(t)}{S(t)}\right] \tag{1.10}$$

If the survival function is known; it is easy to determine hazard function and vice versa, for more details see Lee et al.[34] and Hosmer[35].

1.6.3 Kaplan-Meier Survival Curves and the Log Rank Test

To estimate and plot survival curves and to compare two or more survival curves defined by PFs, a common approach is the Kaplan- Meier (KM) method followed by a log rank test.

Kaplan-Meier Method

The Kaplan-Meier estimator[32 35 36] (also known as the product limit estimator) obtains a non-parametric estimate of the survival function from the data. Suppose that a sample of n patients are obtained from a population with survival function $S(t)$ and $t_1 < t_2 < \cdots < t_k$ are the observed event times (e.g. death) where $k \leq n$, as some patients may be censored.

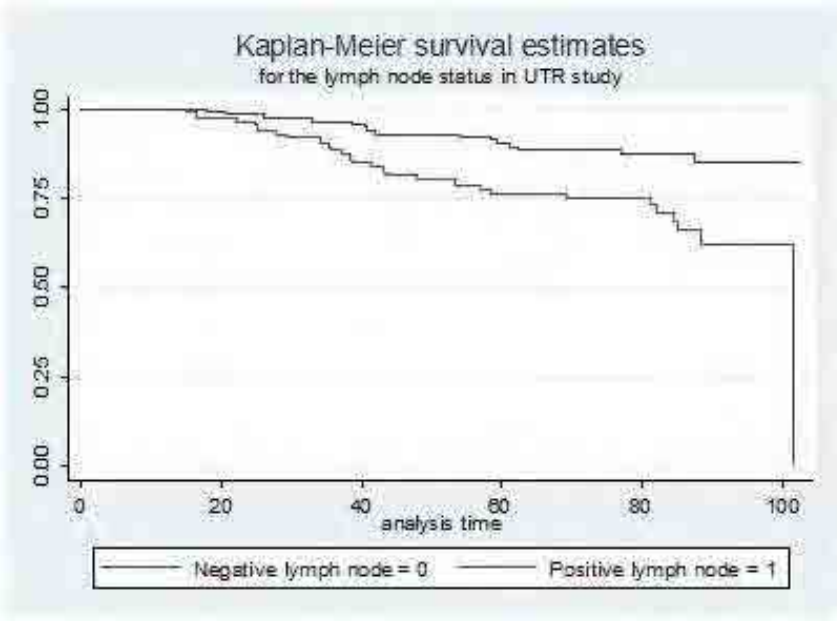
Let n_i be the number of observations at risk at the time prior to t_i. let d_i be the number of events occurs at time t_i. The survival function $S(t)$ can be estimated by using Kaplan-Meier formula which can be written as follows:

$$\hat{S}(t) = \prod_{t_{i<t}} \left(\frac{n_i - d_i}{d_i} \right) \tag{1.11}$$

As an example consider the 'Utr' study selected from the IPD of Look et al.[14] , see Table 1.2; this study has 199 subjects and 42 died. This dataset is divided into two groups defined by a potential PF called Lymph node status that is either positive or negative; the aim is to assess whether the Lymph node status has influence on the breast cancer patients.

The Kaplan-Meier graph shown in Figure 1.8 is obtained from survival estimates from applying to equation 1.11 to the data in Table 1.2. In Figure 1.8 the estimated survival function for the patients with negative lymph node lies completely above that for the patients with positive lymph node, which indicates that the patients who have negative lymph node have a more favourable survival experience (live longer). To test if there is a statistically significant difference between the two groups, the log rank test can be used.

Figure 1.8 Kaplan Meier curve for the positive lymph node group



The log rank test is a large sample chi-square test, and tests the null hypothesis that the curves are equal. It calculates the expected and observed numbers of events at each event time for each group. For more details, see Collett [37]. The results of the log rank test for the lymph node status are illustrated in Table 1.5. It gives a P-value of < 0.01. This indicates statistically significant evidence that there is a difference between the survival curve for positive and negative Lymph node status, thus lymph node status appears to be a PF in breast cancer.

Table 1. 5: Log rank test results and P-value for the equality of positive and negative lymph node status for all Utr study

Lymph node status	**Observed deaths**	**Expected death**
Negative lymph node	15.00	25.44
Positive lymph node	27.00	16.56
Total	42.00	42.00
		P-value <0.001
		chi2(1)=10.91

As a log rank test is a hypothesis test, it does not estimate the difference in survival or the difference in event risk between groups, but this can be addressed by calculating a hazard ratio and using Cox proportional hazard models, as now described.

1.6.4 Hazard Ratio

The hazard ratio is the effect of an explanatory variable (PF) on the hazard or risk of an event. It compares the (instantaneous) risk of an event in one group relative to the risk in another group[38] [39]. An estimate of the hazard ratio can be obtained from the log rank test as follows:

$$HR = \frac{O_1/E_1}{O_2/E_2} \tag{1.12}$$

Where O_1 and O_2 are the observed number of the event in each group; E_1 and E_2 are the expected number of events. Applying to our example, the hazard ratio can be calculated from Table 1.5 as follows

$$HR = \frac{15/25.44}{27/16.56} = \frac{0.5896}{1.6304} = 0.361 \tag{1.13}$$

The hazard ratio of 0.361 indicates that the risk of death into negative lymph node group is 0.361 times (36% of the risk in the positive lymph node group).

The confidence interval for the hazard ratio can be calculated as follows:

$$\exp[\ln(HR) \pm 1.96 \times s.e.(\ln(HR))] \tag{1.14}$$

where:

$$s.e.\ln(HR) = \sqrt{\frac{1}{E_1} + \frac{1}{E_2}} \tag{1.15}$$

Applying to our example the confidence interval for hazard ratio is calculated as follows:

$$s.e.\ln(HR) = \sqrt{\frac{1}{25.44} + \frac{1}{16.56}} = \sqrt{0.099} = 0.315 \tag{1.16}$$

$$exp[\ln(0.361) \pm 1.96 \times (0.315)] \tag{1.17}$$

So we are 95% confident that the hazard ratio lies between 0.20 and 0.67, which does not include the null value of 1 which confirms the significant result that we obtained from the log rank test.

1.6.5 Cox Proportional Hazards Model

A more flexible way to estimate the hazard ratio (HR) that allows continuous multiple PFs to be considered is a Cox proportional hazard model, which is the most commonly model for analysing survival data. It is expressed as follows:

$$\lambda(t) = \lambda_0\,(t)\, e^{\sum_{i=1}^{k} \beta_i\, x_i} \tag{1.18}$$

Where $\lambda(t)$ the hazard is at time t, x_i is an explanatory variable (i.e. a PF), k indicates the number of the variables, and $\lambda_0\,(t)$ is the baseline hazard for patients with x_i values equal to zero.

The Cox proportional hazard model and indeed the log-rank test assume that the hazard ratio is constant over time. In other words, suppose that there are two hazards for two patients who differ in just x_1 by one unit, and then the hazard ratio can be calculated as:

$$\widehat{HR} = \frac{\lambda(t, X_1 + 1)}{\lambda(t, X_1)} = \frac{\lambda_0\ (t) exp\left[\hat{\beta}_1 (X_1 + 1)\right]}{\lambda_0(t) exp\left[\hat{\beta}_1\ X_1\right]} = exp\ [\hat{\beta}_1\ ] \tag{1.19}$$

From the previous equation, it can be seen that the baseline hazard function for the two patients is cancelled out of the formula and the net result does not contain t, thus the hazard ratio is a constant over time.

The Cox proportional hazard model is a semi-parametric model, this means that the $\lambda_0(t)$ is an unknown function and is left unspecified. The model can be estimated by using maximum likelihood estimation of the partial likelihood to give parameter estimates denoted as $\widehat{\beta}_i$, see Hosmer et al[35]. The partial likelihood is given as:

$$l_p\ (\beta) = \prod_{i=1}^{n} \left[\frac{\exp[X_i'\ \hat{\beta}]}{\sum_{j \in R(t_i)} \exp[X_i'\ \hat{\beta}]} \right]^{c_i} \tag{1.20}$$

where the summation in the denominator is over all subjects in the risk set at time t_i, denoted by $R(t_i)$, $c_i = 1$ for uncensored observations, $c_i = 0$ for censored observations, k is the number of PF's for subject i denoted by the vector $X'_i = (x_{i1}, x_{i2}, x_{i3}, \ldots\ldots, x_{ik})$. This vector could include any types of PF's (e.g. continuous PF, dichotomized PF). By taking the logarithm for equation 1.20, the value of $\hat{\beta}$ can be found by taking the first derivative with respect to β, equalling the equation to 0 and solving it; the variance of the $\hat{\beta}$ can be calculated by the same manner, by taking the inverse of the negative of the second derivative of the log partial

likelihood with respect to β and equalling the equation with 0 and solving it, for more detail see Homser et al.[35]

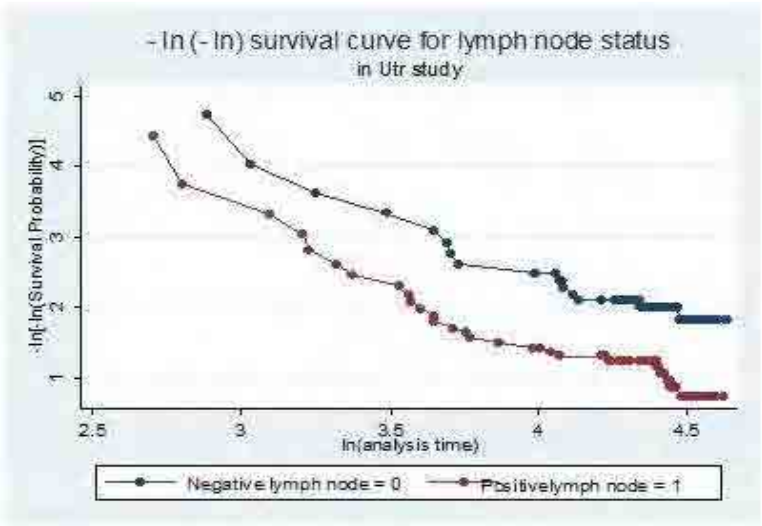
1.6.6 Test of Proportional Hazard Assumption

The assumption of Cox proportional hazard model is that the hazard ratio is constant overtime. There are many methods to test the assumption of Cox proportional hazard model[34]. The two common of these methods are as follows:

$-Ln-Ln$ Survival Curve

This is a graphical technique which allows comparing estimated $(-ln(-ln))$ survivor curves over different categories of PFs. Parallel curves, say comparing positive with negative lymph nodes status, indicate that the Cox proportional hazard assumption is satisfied, for more details, see Kleinbaum et al.[40]. Applying to our example for 'Utr' study, the $(-ln(-ln))$ survival curve for positive and negative lymph nodes for 199 patients is shown as follows:

Figure 1.9: $-ln(-ln)$ survival curve for lymph node status for all subjects in Utr study



It can be seen from Figure 1.9 that the assumption of Cox proportional hazard model appears reasonable because the two curves are parallel. Note that $-ln(-ln)$ survival curve is only

valid to check the Cox regression assumption for the categorized PF; in particular, it cannot be applied with continuous scale PF. The approach of goodness of fit test (e.g. Schoenfeld residuals) can be used to check the proportional hazard assumptions for the continuous PF (see below).

The Goodness of Fit Test Approach

The most common test uses Schoenfeld residuals, which is proposed by Shoenfeld[41]. For each PF in the model, Schoenfeld residuals are defined for every subject who has an event. The estimator of the Schoenfeld residuals for the i^{th} subject on the k^{th} PF is given as

$$\widehat{r_{ik}} = c_i \, (x_{ik} - \bar{\hat{x}}_{w_{ik}}) \tag{1.21}$$

where

$$\bar{\hat{x}}_{w_{ik}} = \frac{\sum_{j \in R(t_i)} x_{jk} \;\; \exp[\acute{X}_j \, \hat{\beta}]}{\sum_{j \in R(t_i)} \exp[\acute{X_j} \hat{\beta}]} \tag{1.22}$$

Grambsh and Therneau[42] suggest scaling the Schoenfeld residuals by an estimator of the covariance matrix of the estimated coefficients by the number of events (i.e. the observed number of uncensored survival times m), yields residuals with greater diagnostic power than the unscaled residuals. Thus, the approximate scaled Schoenfeld residuals[43] are given as :

$$\hat{r}^*_{ik} = m \, \text{Var} \, (\hat{\beta}) \hat{r}_{ik} \tag{1.23}$$

where $\hat{r}_{ik}$ can be obtained from equations 1.21 and $\text{Var}(\hat{\beta})$ can be calculated from equation 1.20 as mentioned before.

Consider a Cox proportional hazard model with two PFs: uPA and lymph node status. Then, there are two Schoenfeld residuals defined for each subject who has an event, one for each of the two PFs. If the proportional hazard assumption holds for a particular PF, then the Schoenfeld residuals for that PF will not be related to survival time. The P-value is used to decide whether

the proportional hazard assumption holds. If the P-value is less than certain values (say 0.10), this provides evidence that the proportional hazard assumption does not hold. Furthermore, the graphs of Schoenfeld residuals against survival time can be used to check the adequacy of the proportional hazards model. The presence of certain patterns in these graphs may indicate departures from the proportional hazards assumption, while extreme departures from the main cluster indicate possible outliers or potential stability problems of the model (see below).

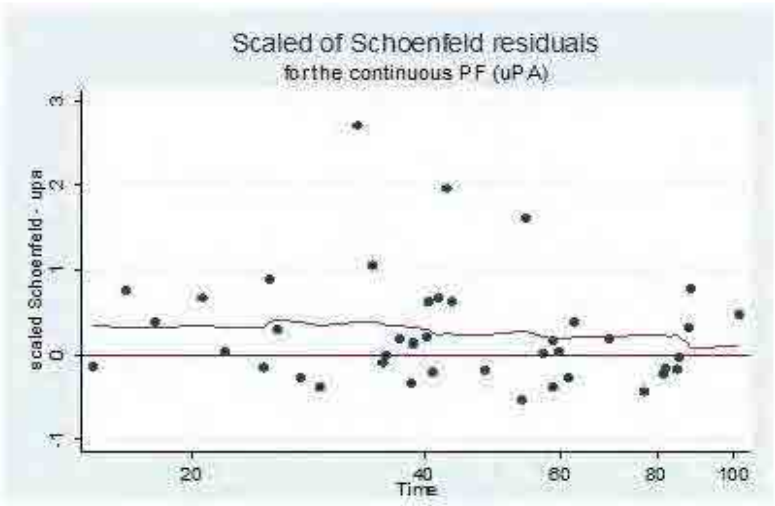
1.6.7 Example of Cox Regression Analysis with Multiple PFs

To illustrate the use of Cox regression model for multiple PFs, two PFs within Look et al.[14] were selected from the 'Utr' study (uPA and lymph node status) as mentioned before; uPA is a continuous variable and lymph node status is a dichotomous variable (positive hormone receptor =1 and negative hormone receptor = 0). A Cox proportional hazard model is fitted as follows:

$$\lambda(t) = \lambda_0(t)exp[\beta_1 \cdot \text{uPA} + \beta_2 \cdot \text{lymphnode}] \qquad (1.24)$$

The goodness of fit test using Schoenfeld residuals was used to check the assumption of the proportional hazard model which gave P-value of 0.32 and 0.95 for the uPA and lymph node factors respectively. Note that each of those P-values test the assumption for one variable given that the other PFs are included in the model. For example, the P-value for uPA (0.32) assesses the proportional hazard assumption for uPA, assuming the proportional hazard assumption is satisfied for lymphnode status. As the P-value for the two PFs is greater than a certain value (0.10), this indicates no significant evidence against the proportional hazard model assumptions. Figure 1.10 and Figure 1.11 illustrate the scaled Schoenfeld residuals for uPA and lymph nodes PFs as follows:

Figure 1.10: The scaled Schoenfeld residuals for the uPA



It can be seen from Figure 1.10 that the horizontal axis represents the time and the vertical axis represents the scale Schoenfeld residuals for uPA. The assumption of the Cox proportional hazard model is acheived for the uPA as a continuous variable; this is because there is no a strong trend between the Schonfield residuals and the original time.

Figure 1.11: The scaled Schoenfeld residuals for the lymph node

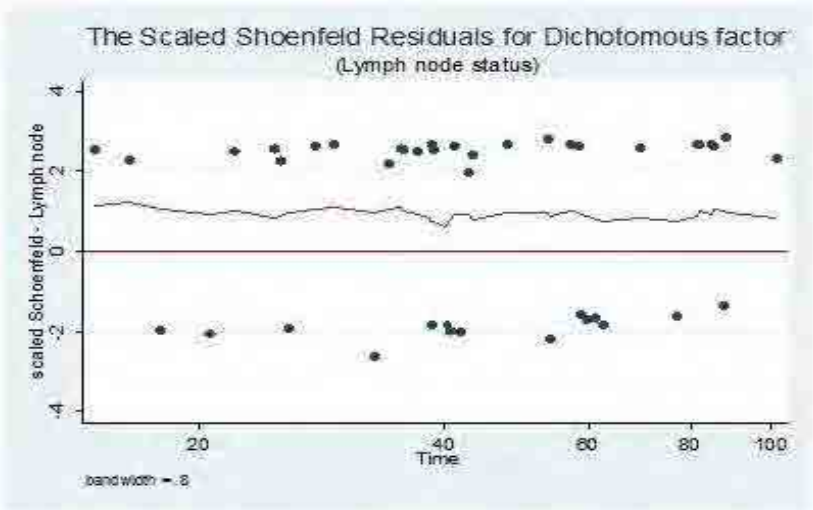


Figure 1.11 shows the scaled Schoenfeld residuals for the lymph node status as a binary PF. There are two bands of residuals for the dichotomous lymph node PF. The upper band corresponds to subjects with positive lymph node (np=1) and the bottom one to those with negative lymph node (np=0). As can be seen from the graph, there is no a strong trend between

the scaled schonfeld for the lymph node status as a PF and the original time, whether when the lymph node status is positive or negitave.

After verifying the proportional hazard model assumption, the Cox proportional hazard model estimates are shown in Table 1.6. Both uPA and lymph node appear to be PFs. For example, the hazard ratio for lymph node status means that the patients with positive lymph node have a 2.23 times higher death risk than those with negative lymph node status throughout the study period. Also, the confidence interval for the hazard ratio for lymph node PF does not include one which means that the hazard ratio is statistically significant and this is confirmed by the P-value which is 0.006.

Table 1. 6: Cox proportional hazard estimates for all subjects in Utr study

Variables	**uPA**	**Positive Lymph nodes**
Hazard Ratio (HR)	1.21	2.23
Standard Error of ln(HR)	0.12	0.65
Confidence Interval (95%)	1.01-1.46	1.27-3.94
P-value	0.05	0.006

1.7 Systematic Reviews and Meta-analysis of Prognostic Factor Studies

1.7.1 Systematic Reviews

Systematic reviews are considered as an essential tool in health care research because they greatly facilitate evidence-based clinical practice and healthcare policies. Systematic reviews seek to identify, collate and appropriately summarise the existing evidence on a particular topic from published and unpublished existing research studies. The whole process is done systematically, so that the review process is transparent and reproducible. The review allows overall evidence-based conclusions to be formed and helps to identify the questions to be

addressed in future primary studies. They are practically crucial for identifying PFs, as numerous research studies exist for each PF and many have conflicting results. Evidence based guidelines for each PF are highly desirable, and indeed essential.

Numerous systematic reviews of PFs exist. For example, in Falagas et al.[44] the authors did a systematic review by searching in Medline and PsycINFO database to examine the effect of psychosocial factors (e.g. social support, marriage, depression and constraint emotions) on the survival of breast cancer patients. They identified 31 studies examining the association of various psychosocial parameters with overall breast cancer survival. They found 25 of the 31 studies showed a statistically significant association between at least one psychosocial variable and disease outcome and 6 studies examining whether psychological intervention influences the disease outcome. In particular, they found that social support, marriage have a positive effect on breast cancer patients; while depression and constraint of emotions have a negative effect on the breast cancer survival.

1.7.2 Meta-Analysis using Aggregate Data

Meta-analysis is a statistical analysis that combines results from multiple individual studies on the same topic. It is often applied as part of the systematic review to provide a quantitative summary of the evidence[45]. Typically meta-analysis seeks to estimate the average value of an effect of interest across studies, such as the prognostic effect of a particular factor. To do this, it is common to extract and combine relevant aggregated data from the identified studies. For a meta-analysis of PFs the choice of aggregated data are usually the odds ratio or the hazard ratio, alongside some measure of their uncertainty (e.g. standard error or confidence interval). These are combined across studies, using an appropriate meta-analysis model, to estimate the average effect size across studies.

Fixed-effect Approach

A fixed-effect meta-analysis assumes that there is no heterogeneity in the effect size across studies[46]. In other words, it means that the true effect size in all studies is the same. For PF studies this means the true PF effect (i.e. hazard ratio or odds ratio) is the same in each study. The fixed-effect model can be written mathematically as follows:

$$\hat{\beta}_i = \beta + e_i$$
$$e_i \sim N(0, S_i^2)$$

where $\hat{\beta}_i$ is the effect estimate in study i (e.g. log hazard ratio), S_i^2 is the known variance of $\hat{\beta}_i$ (assumed known), β is the fixed effect size, e_i represents the sampling error for each study and $i = 1,2,\ldots,k$ represents the number of studies to be combined. The main goal of the meta-analysis is estimating the fixed-effect size β. One method to achieve this is the inverse variance method, where each study estimate is given a weight inversely proportional to S_i^2. Thus, the parameter β which represents the pooled effect size of the studies, is estimated as:

$$\widehat{\beta} = \frac{\sum_{i=1}^{k} w_i \widehat{\beta}_i}{\sum_{i=1}^{k} w_i} \quad (1.25)$$

Where $w_i = \frac{1}{S_i^2}$ refers to the weight for each study. This is the maximum likelihood estimate which minimizes the variance of β. The variance of $\hat{\beta}$ is found by (Borenstein et al.[46]):

$$Var(\hat{\beta}) = \frac{1}{\sum_{i=1}^{k} w_i} \quad (1.26)$$

The confidence interval for $\hat{\beta}$ can be written as:

$$\widehat{\beta} \pm 1.96 \times \sqrt{Var(\hat{\beta})} \quad (1.27)$$

For ratio measures such as the hazard or odds, it is best to apply the meta-analysis on the natural log scale as this will be more symmetric and approximately normal; to convert back to the original scale one can take exponential of $\hat{\beta}$ and the exponential of the confidence interval in 1.25 and 1.27.

Random-effects Approach

In the medical literature, often there is no heterogeneity among studies. Studies often differ in their design, clinical and patients characteristics, methods, analysis...etc. This may cause the effect of interest to vary across studies. As a result a random-effects meta-analysis method will be more approachable, as this estimate show the true effect size vary across studies. Let the notation be as before, with additionally Θ_i the true effect size in the i^{th} study. The random-effects model can be written as follows

$$\begin{aligned} \hat{\beta}_i &= \Theta_i + e_i \\ &= \beta + u_i + e_i \end{aligned} \tag{1.28}$$

where $u_i \sim N(0, \tau^2)$ and $e_i \sim N(0, S_i^2)$. In particular, u_i represents the difference between the study specific effect size Θ_i and the average study effect size β; e_i is the sampling error in each study, with known variance S_i^2, and τ^2 is the between study variance. For more clarification, the total variance of each $\hat{\beta}_i$ around β can be written as:

$$Var(\hat{\beta}_i) = \tau^2 + S_i^2 \tag{1.29}$$

It can be seen from equation (1.29) if the between study variance is zero, then the random-effects model will reduce to the fixed-effect model.

There are a number of different methods to estimate the random-effects model, for which both β and τ^2 must be estimated. For example, restricted maximum likelihood can be used to iteratively

estimate both τ^2 and β simultaneously. Else a method of moment estimate of τ^2 can be obtained, using the Dersimonian and Laird method[46] :

$$\hat{\tau}^2 = \frac{Q-(k-1)}{U} \tag{1.30}$$

where Q is the value of heterogeneity test statistic, calculated as $Q = \sum_{i=1}^{k} w_i\beta_i^2 - \frac{(\sum_{i=1}^{k} w_i\beta_i)^2}{\sum_{i=1}^{k} w_i}$, k is the number of studies, $U = (k-1)(\overline{w} - \frac{{S_w}^2}{k\overline{w}})$, where $\overline{w}$ and S_w^2 are the mean and variance of the weights from the k studies respectively, $\overline{w} = \sum_{i=1}^{n} \frac{w_i}{k}$ and $S_w^2 = \frac{1}{k-1}(\sum_{i=1}^{k} w_i^2 - k\overline{w}^2)$, then the pooled value $\hat{\beta}$ and its variance can then be estimated

by

$$\hat{\beta} = \frac{\sum_{i=1}^{k} \hat{\beta}_i w_i^*}{\sum_{i=1}^{k} w_i^*} \tag{1.31}$$

The maximum likelihood solution with the study weights $w_i^* = \frac{1}{S_i^2+\tau^2}$

The variance is estimated as $Var(\hat{\beta}) = \frac{1}{\sum_{i=1}^{k} w_i{}^*}$ and confidence interval as in equation 1.27. The random-effects confidence interval will be wider than the confidence interval for the fixed-effect method because the between study variance is accounted for if $\tau^2 > 0$.

The random-effects method is usually more preferable for PFs studies, as heterogeneity usually exists. PF studies often differ in quality, types of participants, methods of measurement, and choice of statistical techniques (e.g. cut offs, analysis methods, length of follow up, adjustment factors) amongst other factors that create heterogeneity.

The I^2 statistic is a tool for measuring heterogeneity. The I^2 statistic assess the extent of heterogeneity in a meta-analysis. Higgins et al.[47] interpreted I^2, as percentage of the total variability in a set of effect sizes due to between study heterogeneity. For example $I^2 = 50\%$

means that half of the total variability among effect sizes is caused by between study heterogeneity .

Higgins and Thompson [47] tentatively proposed a guideline for interpertating the magnitude of I^2 as 25% means low heterogeneity, 50% means medium heterogeneity, and 75% means high heterogeneity.

Note that under the fixed-effect model, as mentioned before, the true effect size for all studies is assumed to be identical, and the only source of uncertainty is the within study error (sampling error). By contrast, under the random-effects model, the true effect size across studies is different, because there are two sources of variation. So, in a random-effects model of PFs, β gives the mean effect size from the distribution of effect sizes across studies. Traditionally, random-effects meta-analysis focuses on the summary effect size β and its confidence interval. But this says nothing about how the true effects in each study are distributed about the mean effect. To address this, Higgins et al.[47] suggest calculating a 95% prediction interval for the effect in new study as follows:

$$\hat{\beta} \pm 1.96\sqrt{\hat{\tau}^2 + Var(\hat{\beta})} \qquad (1.32)$$

where $\hat{\beta}$, as mentioned before, is the mean effect size, $\hat{\tau}^2$ is the estimate of the variance of the true effect sizes, and $Var(\hat{\beta})$ is the variance of $\hat{\beta}$.

1.7.3 Meta-Analysis Examples

As mentioned, Hukkelhoven et al.[30] assess the effect of age as a PF for severe traumatic brain injury and they obtained IPD data for four trials.

Table 1. 7: Odds ratio and standard error estimates for logistic regression for the four trials

Trial number	Odds ratio	Coefficient (Ln (Odds ratio))	Standard error of ln (Odds ratio)
Trial 1	1.364	0.310	0.064
Trial 2	1.302	0.262	0.050
Trial 3	1.456	0.375	0.053
Trial 4	1.388	0.328	0.085

Table 1.7 gives the odds ratio estimates for each study, for two patients who differ in age by 10 years and the standard error of the ln (odds ratio) as obtained from a logistic regression model including age as a linear prognostic effect in each study separately (see equation 1.6).

Table 1. 8: Meta-analysis estimates for fixed-effect and random-effects meta-analysis applied to the TBI data of Hukkelhaven et al. assessing whether age is a PF of 6 month mortality

	Ln odds ratio (OR)	95% CI for (Ln OR)	OR	95% CI for OR
Fixed-effect	0.32	$0.26 - 0.37$	1.37	$1.29 - 1.45$
Random-effects	0.32	$0.26 - 0.37$	1.37	$1.29 - 1.45$

$p < 0.01$ (for the model)
$I^2 = 0$ and $p = 0.492$ (from Q test for heterogeneity)

Table 1.8 gives the parameter estimates when fitting a fixed and random-effects model, using maximum likelihood estimator using the STATA 'metan' command[48] [49]. As there is no heterogeneity between studies ($I^2 = 0\%$), the fixed-effect estimate is the same as random-effects estimate (1.37), which means that the odds of mortality by 6-months increases by 1.37 times as a result of increasing age by 10 years. The confidence interval does not include one, which indicates that there is a significant relationship between age and outcome (6 month mortality), i.e. age is a PF.

Figure 1.12: The forest plot for fixed-effect of meta-analysis of the four TBI studies

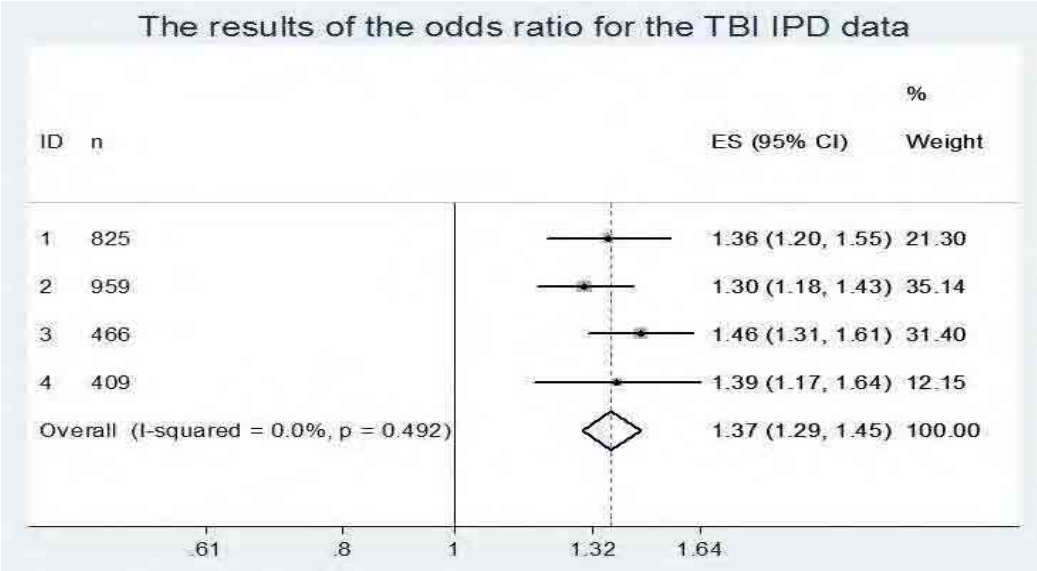


Figure 1.12 gives the forest plot for this meta-analysis. The black dots represent the study effect estimates and the horizontal line represent their associated 95% confidence interval. The centre and width of the diamond give the estimate and 95% confidence interval of the odds ratio respectively. The homogeneity in this example is unusual for PF meta-analysis, but perhaps arises from the four TBI studies all being trials. Usually, PF studies are observational studies that are less well designed and will have heterogeneity. An example of this for the Look et al.[14] is shown in detail in chapter 3.

1.7.4 The Problems of Meta-analysis of Prognostic Factor Studies using Aggregate Data

Though meta-analysis of PF studies are clearly desirable, unfortunately there are many problems for the approach, which are briefly described below

Difficulty of Identifying Relevant Publications

Due to a lack of taxonomy, it is difficult to search for PF studies. PFs are termed prognostic markers, predictive markers, and prognostic variables, amongst many other names[45]. Further,

individual PFs often have multiple names themselves, like NMYCN, N-MYC, MYCN in neurosblastoma[50]. There are currently no standards for conducting a literature review and search of PF studies.

Publication Bias and Other Sources of Bias

Publication bias is considered as one of the biggest problems in PF studies. It occurs when the results of PF studies do not give statistically significant or clinically valuable findings, and so are not published. Furthermore, other dissemination bias problems may exist such as: (i) non-statistically significant or negative results may not be reported in as much detail as significant results[51]; (ii) there may be outcome and sub-group reporting bias[22]; (iii) language bias, where non-english studies are only written-up in English if they are significant[22] and (iv) the biased choice of optimum cutoff point (for dichotomizing continuous PFs) which researchers often choose to produce the most significant result[8].

Poor Primary Study Design

Currently, many published primary studies do not have a protocol, are not designed properly, and are of poor quality. For instance, there is no prior specification of the PFs to be investigated; no sample size calculation; and no adjustment for other PFs. Also, due to poor reporting; it is often hard to distinguish good quality and bad quality PF studies[25].

Inadequate Reporting of the Methods

The general standard of reporting primary PF studies are inadequate[52 53]. For example, there is often insufficient information reported about the statistical analysis (e.g. what is the statistical model that used, how the assumptions were checked); basic information such as the number of patients and events in the groups defined by marker level are often not provided; the inclusion and exclusion criteria may not be clear or given and results such as hazard ratio, confidence

interval and P-value are often missing or partially reported. As a consequence, not all of the available PF studies can be incorporate in a meta-analysis of aggregate data, which may cause bias and inefficiency.

Heterogeneity

Heterogeneity is one of the most common problems in meta-analysis of PF studies. For example, there is often great diversity in the choice of the cut-off points that used to dichotomize the continuous PFs (e.g. age) across studies; the choice of adjustment factors; the stage of disease; the type of treatments; the methods of measurement; among many other factors. Heterogeneity across studies can limit the interpretation of meta-analysis results, especially when the direction of PF effect appears inconsistent across studies[46].

For such reasons the aggregate data approach to meta-analysis of PFs has been criticized and often only serves to show problems. This has been shown in numerous existing reviews of PFs [54]and discussed in many commentaries[55]. To help overcome the issues, an alternative IPD approach has been championed[56-59].

1.8 Individual Patient Data versus Aggregated Data

An IPD approach to meta-analysis utilizes the raw data for each individual study, rather than the available aggregated data. Example IPD for PF meta-analysis was shown in Table (1.1) and (1.2). IPD can potentially overcome many of the problems discussed in section 1.7.4; for example, it allows one to[11 56 60-63].

- Use consistent inclusion and exclusion criteria across studies, and if appropriate reinstate individuals into the analysis who were originally excluded. For example in Look et al. dataset one could study uPA after adjusted or grouped by hormone receptor for all 15 studies.

- Observe and account for missing data at the individual-level.
- Verify results presented in the original study publications (assuming IPD provided can be matched to that IPD used in the original analyses).
- Use up-to-date follow-up information, which is potentially longer than that used in the original study publications.
- Identify those studies which contain the same or overlapping sets of participants.
- Calculate and incorporate results for those missing or poorly reported outcomes and summary statistics across published studies; it may thus reduce the problem of selective within-study reporting.
- Calculate and incorporate results for unpublished studies; it may thus reduce the problem of publication bias.
- Standardize the strategy of statistical analysis across studies (e.g. the analysis method, how continuous variables are analysed, the time-points assessed etc) and use more appropriate/advanced methods where necessary.
- Assess model assumptions in each study, such as proportional hazards in Cox regression models, and model complex relationships like time-dependent effects.
- Produce estimates adjusted for existing PF, where previously only unadjusted estimates were available; this may increase statistical power and allow the independent prognostic ability of a factor to be identified.
- Adjust for a consistent set of PFs across studies.
- Obtain meta-analysis results for specific subgroups of participants across studies (e.g. those receiving a particular treatment, those with a particular stage of disease), and assess differential PF effects across individuals; this facilitates individualized or stratified estimates, and can help reduce observed between-study heterogeneity.

An IPD meta-analysis is not without its disadvantages[11 55 64]. In particular, it is known to be resource intensive, as substantial time and costs are required to contact study authors; obtain their IPD; input and 'clean' the provided IPD; resolve any data issues through dialog with the data providers; and generate a consistent IPD format across studies. For example, in 2002 the IPD meta-analysis of Ioannidis et al.[65] required 2,088 hours for data management, with 1000 e-mails exchanged between study collaborators and the data managers. The required cost and time will clearly vary depending on the complexity and number of studies involved[10]. But such factors need serious consideration before embarking on an IPD meta-analysis or when applying for grant income. In particular, resource requirements must be considered for both the team conducting the IPD meta-analysis and the original study authors themselves; the latter are often neglected but they are crucial to the success of the project and will often commit many hours 'cleaning' and updating their data, and resolving ongoing queries.

1.9 Statistical Methods for IPD Meta-Analysis

As for any meta-analysis, an IPD meta-analysis aims to summarise the evidence from multiple related studies regarding a particular clinical question, such as whether a factor is prognostic. It seems sensible the statistical implementation of an IPD meta-analysis to preserve the clustering of patients within studies, and not analyse the IPD as if all coming from a single study (although this has not been explicitly shown, see chapter 4 and 5). This can be achieved using a 'one-step' or a 'two-step' approach[10]. In the 'two-step' approach, the IPD are first analysed separately in each study using an appropriate statistical method for the type of data being analysed; for example, for binary outcome a logistic regression model might be fitted, or for time-to-event data a Cox regression might be applied. This produces aggregate data for each study, such as the PF effect estimate and its standard error; these are then synthesized in the second step using a

suitable model for meta-analysis of aggregate data, such as one weighting by the inverse of the variance whilst assuming fixed or random PF effects across studies (see section 1.7).

In the 'one-step' approach, the IPD from all studies are modeled simultaneously whilst accounting for the clustering of participants within studies; this again requires a model specific to the type of data being synthesized, alongside appropriate specification of the meta-analysis assumptions (e.g. fixed or random-effects across studies). Detailed statistical articles regarding the implementation and merits of 'one-step' and 'two-step' IPD meta-analysis methods are available; for example, see Turner et al. [66], Higgins et al.[67] and Whitehead et al.[68] . The approaches have been shown to give very similar results, particularly when interest lies in a single treatment effect estimate[69], though this has not been considered in much details for PFs (see chapter 4 and 5).

One-step IPD meta-analyses more conveniently require only a single model to be specified, but this may increase complexity for non-statisticians and requires careful separation of within-study and between-study variability[60]. Two-step IPD meta-analyses are clearly more laborious, but in the second-step they allow traditional, well-known meta-analysis techniques such as those used by the Cochrane Collaboration (e.g. inverse-variance fixed-effect or random-effects approach; Mantel-Haenszel method). Importantly both 'one-step' and 'two-step' IPD meta-analysis approaches produce results to inform evidence-based practice, such as the pooled PF effect across studies and how the PF effect is modified by study-level characteristics (e.g. treatment, study location) and patient-level characteristics (e.g. age, stage of disease). IPD meta-analysis models ('one-step' and 'two-step') will be considered in details in chapters (4), (5) and (6).

1.10 Aims and Outlines of the Thesis

An IPD meta-analysis is clearly a desired approach to evidence synthesis of PF studies and considered the best way to facilitate evidence-based use of PFs in health care. As mentioned, in this thesis I have two large IPD meta-analysis datasets available, which I will use to explore these advantages further. However, IPD does not necessarily overcome all the problems of PF meta-analysis such as poorly designed studies, missing patient data, and different methods of measurements. There has also been little methodological or empirical research of how to conduct an IPD meta-analysis of PFs, and this thesis aims to address this.

My research in this thesis begins in chapter (2) with a systematic review to collate existing IMPF articles and examine a sample of these in detail to identify the current advantages and problems of the IPD approach. This allows me to form an appropriate research agenda to motivate the subsequent research in this thesis to overcome the problems identified. In chapter (3) I then show some real examples of why the IPD approach is preferred to an aggregated data approach from the statistical prospective, using the provided IPD in breast cancer. At the same time, this allows me to illustrate some of the problems identified in chapter (2) that need to be addressed by further research. In chapter (4) I develop possible IPD meta-analysis models according to whether a one-step or two-step meta-analysis approach is used; whether the model accounts for the clustering of patients within studies; and whether the model accounts for the residual variation for the same values of a PF in each study. In Chapter (5) I then use a simulation study to investigate further three models of IPD meta-analysis that have been developed in chapter (4), to investigate and examine whether we should ignore the clustering of patients within-studies and which is the best approach (one-step or two-step). In Chapter (6) I investigate what we should do when the continuous PF does not exhibit a linear prognostic effects; in particular I use a fractional polynomial approach to overcome this problem and fit possible fractional

polynomial IPD meta-analysis logistic regression models by using one-step and two-step approaches. In chapter (7) I consider the potential problem of small-study effects and availability bias that may arises in IMPF; in particular, I review IMPF articles and extract forest plot statistics to allows small-study effects to be examined using funnel plots, cumulative meta-analysis, statistical tests and adjustment methods. Finally, in chapter 8 I conclude with a summary and discussion of the key findings of this thesis with suggestions for further research.

CHAPTER 2

A REVIEW OF EXISTING META-ANALYSIS OF PROGNOSTIC FACTOR STUDIES USING IPD

2.1 Introduction

As discussed at the end of chapter 1, an IMPF has many potential advantages over the aggregate data approach, but it will have its own challenges and potential pitfalls, and may not solve all the problems of conducting meta-analysis of PFs. In this chapter I will undertake a systematic review to identify IMPF articles in the published medical literature. The aim of this review is to identify how common the IPD approach is; to assess the trend in it application over time; and to critically examine a random sample of IMPF studies; in particular, to examine how an IMPF is conducted (e.g. how IPD are obtained, how statistical analyses are performed etc.) and to evaluate how IMPF projects are reported. The latter is particulary important to establish the current state of the field and to identify common challenges and problems to motivate the remainder of the thesis. The methods and results of the systematic review are described in the next two sections followed by some discussion and the rationale for remaining chapters.

2.2 Methods of the Review

2.2.1 Identification of General IPD Meta-analysis Articles

It was deemed very difficult to identify IMPF articles directly[60], due to the huge variety of disease fields, inconsistent nomenclature (e.g. 'predictive', 'prognostic'), and numerous types of PFs (e.g. genes, tumor markers, chromosomes, biomarkers, clinical characteristics etc). So, to identify IMPF studies, a wider search was initially undertaken for *any* type of IPD meta-analysis articles (i.e. IPD meta-analysis articles in PF and non-PF settings).

An existing database of IPD meta-analysis articles were available from Riley et al.[54]; this contained 199 articles identified from January 1992 to the week 24 of 2005, and these were included in my database. To update this database, I also searched Medline, Embase and the Cochrane Library from the 1st Jan 2005 to March 2009 using the same search strategy as Riley et al.[54] see Table (2.1). I also searched Google crudely using 'Individual Patient Data Meta-analysis', to check for any omissions (Figure 2.2).

Articles must have reported an IPD meta-analysis to be included in my database, that synthesized IPD across multiple studies or collaborative groups. Note that these articles did not need to be part of a systematic review, as often IPD meta-analysis is done on a 'convenient' set of IPD available from collaborators. There was also no restriction on included articles according to type of studies being synthesized (e.g. randomized trials, observational studies,...etc). However, methodological articles that reported an IPD meta-analysis to demonstrate their methods were excluded, as were articles reporting single multi-centre trials. The abstracts of all articles identified by the search were read by me and then classified as either 'yes', 'unsure', or 'no'. Richard Riley double checked for all articles which I classified as 'yes' or 'unsure', and checked 10% of the ‘no’ articles. Any discrepancies were resolved by discussion. A final decision on the 'no abstract' articles was made after obtaining the full paper.

Table 2. 1: The search strategy used to identify articles in Medline, Embase, and the Cochrane Library

1.(individual patient $ adj6 data.ti.ab.)
2.(individual patient $ adj6 report$.ti.ab.)
3.(individual patient $ adj6 outcomes$.ti.ab.)
4.(individual patient $ adj6 levels$.ti.ab.)
5.(individual patient data.ti.ab.)
6.(ipd.ti.ab.)
7.(individual subject$ adj6 data.ti.ab.)
8.(individual subject $ adj6 report$.ti.ab.)
9.(individual subject $ adj6 outcomes$.ti.ab.)
10.(individual subject $ adj6 levels$.ti.ab.)
11.(raw patient $ adj6 data.ti.ab.)
12.(raw patient $ adj6 report$.ti.ab.)
13.(raw patient $ adj6 outcomes$.ti.ab.)
14.(raw patient $ adj6 levels$.ti.ab.)
15.(raw subject$ adj6 data.ti.ab.)
16.(raw subject $ adj6 report$.ti.ab.)
17.(raw subject $ adj6 outcomes$.ti.ab.)
18.(raw subject $ adj6 levels$.ti.ab.)
19.idiopathic.ti.ab.
20.(immediate pigment darkening).ti.ab.
21.(intermittent peritoneal dialysis).ti.ab.
22.(invasive pneumococcal disease).ti.ab.
23.(indirect photometric detection).ti.ab.
24.(interaural phase disparity).ti.ab.
25.or/1–18
26. or/19–24
27.25 not 26

2.2.2 Identification of IPD Meta-analysis of Prognostic Factors (IMPF) Articles

Using the database of all IPD meta-analysis articles, I then read each one to classify articles as IMPF, not IMPF or 'unsure'. Double checking of all classification was again done by Richard Riley, and any discrepancy resolved via discussion.

Recall that - as outlined in chapter 1 - in this thesis I define a PF study to be an assessment of whether one or more factors measured at some point from diagnosis of disease (or onset of some health condition) are associated with some future outcome. Thus, to be relevant in this review, the IPD meta-analysis article must relate to a synthesis of such studies. I call such articles IMPF articles, other IPD articles that were not mainly focused on evaluating PFs were not deemed relevant. For example, studies with a primary objective to assess a treatment effect were excluded, even if they adjusted for PFs. Similarly, so-called predictive factor studies, which mainly focus on identifying factors that predict response of treatment, were also excluded.

Studies that sought only to develop a prognostic model (and not to evaluate PFs themselves but rather the performance of the model) were also excluded. If studies sought to establish the prognostic effect of one or more factors, and also later developed a prognostic model, they were included. Note that PF studies relate to outcomes in patients with an existing disease, or health condition, (see Hemingway et al.[2]) whereas aetiological studies start with patients without disease and look to identify risk factors for onset of disease. IPD meta-analysis of such aetiological studies were excluded, though a record kept of how many such articles existed.

2.2.3 In-depth Evaluation of Recent IMPF

Following the identification of IMPF articles, I performed an in-depth evaluation of those IMPF articles published from 2006 to 2009 inclusive. A data extraction form was developed, which included 58 questions covering five aspects of the rationale, conduct, analysis, design and reporting on the IMPF. Figure (2.1) gives a brief summary of these questions, for the full list of the questions, see Appendix A.

I also catalogued all the issues noted within each article that hindered the IMPF to generate priority research ideas for the rest of this thesis.

Figure 2.1: Summary of the data extraction from involving 58 questions, which was used to extract information about the 20 IMPF articles examined in details

Aims and initiation: This section investigates the aims for the IMPF articles, the location for the first author and whether there was mention of a project protocol and ethics approval.

Process of obtaining IPD: This section examines how researchers identified relevant primary studies (e.g. systematic review, coalition of research groups); how they decided which studies to seek IPD from; the process of obtaining IPD; and problems encountered.

Details of IPD obtained: this section points out to the proportion of studies providing IPD; the total number of patients in the IPD; whether the number of patients and events were reported for each IPD study; whether there was any missing data problems; and whether there was variability in how prognostic factors were measured.

Type and quality of IPD studies: this section highlights the information about the design (e.g. cohort, randomised trials) of studies providing IPD; whether they were published or unpublished; and whether they were assessed for their quality and, if so, how.

Statistical methods used: this section examines whether a statistical methods section was provided; the statistical models used in the meta-analysis (e.g. Cox regression, logistic regression); and how some specific statistical issues were addressed (such as clustering of patients within studies; between-study heterogeneity in PF effects; and the analysis of continuous PFs).

Assessment of publication bias and availability bias: this section highlights if and how researchers examined the potential impact of publication bias (studies unpublished due to non-significant prognostic results) or availability bias (studies providing IPD are a biased portion of the studies from which IPD was desired) in their meta-analysis.

Adherence to reporting guidelines: As a crude measure of adherence to reporting guidelines for meta-analysis, we recorded how many of the articles referenced the reporting guidelines of either MOOSE[70] or QUORUM[71].

Limitations and challenges of an IMPF: We catalogued all the problems that hindered the IMPF approach as reported by the researchers

2.3 Results

This section documents the findings of the review in three stages: the first stage summarizes the classification of *any* type of IPD meta-analysis and looks at their trend over time; the second stage summarizes the classification of IMPF articles and assesses their trend over time; and the third stage describes the findings from the in-depth examination of IMPF articles from 2006 to 2009.

2.3.1 IPD Meta-analysis Articles

After duplicates were removed, the literature search identified 1420 potential IPD meta-analysis found from 1st Jan 2005 to March 2009. I read the abstract of these articles. These were classified and double checked, see Figure (2.2), resulting in 203 deemed an IPD meta-analysis. There were added to those 199 from Riley et al., [54] with 20 further duplicates then removed; followed by adding another 2 articles after checking references. Finally a further one article was identified from the Google search, giving a final total of 385 IPD meta-analysis articles published between 1991 and March 2009.

Figure 2.3 shows the number of applied IPD meta-analysis articles published per year. It can be seen from the graph that it covers the years from 1991 to 2008 and the number of applied IPD meta-analysis articles per year is generally increasing over time. There are just a few articles per year in the early 1990's, but about 50 per year by 2009. This show the IPD approach is achievable and increasingly common.

Figure 2.2: Details of the search and classification of IPD articles

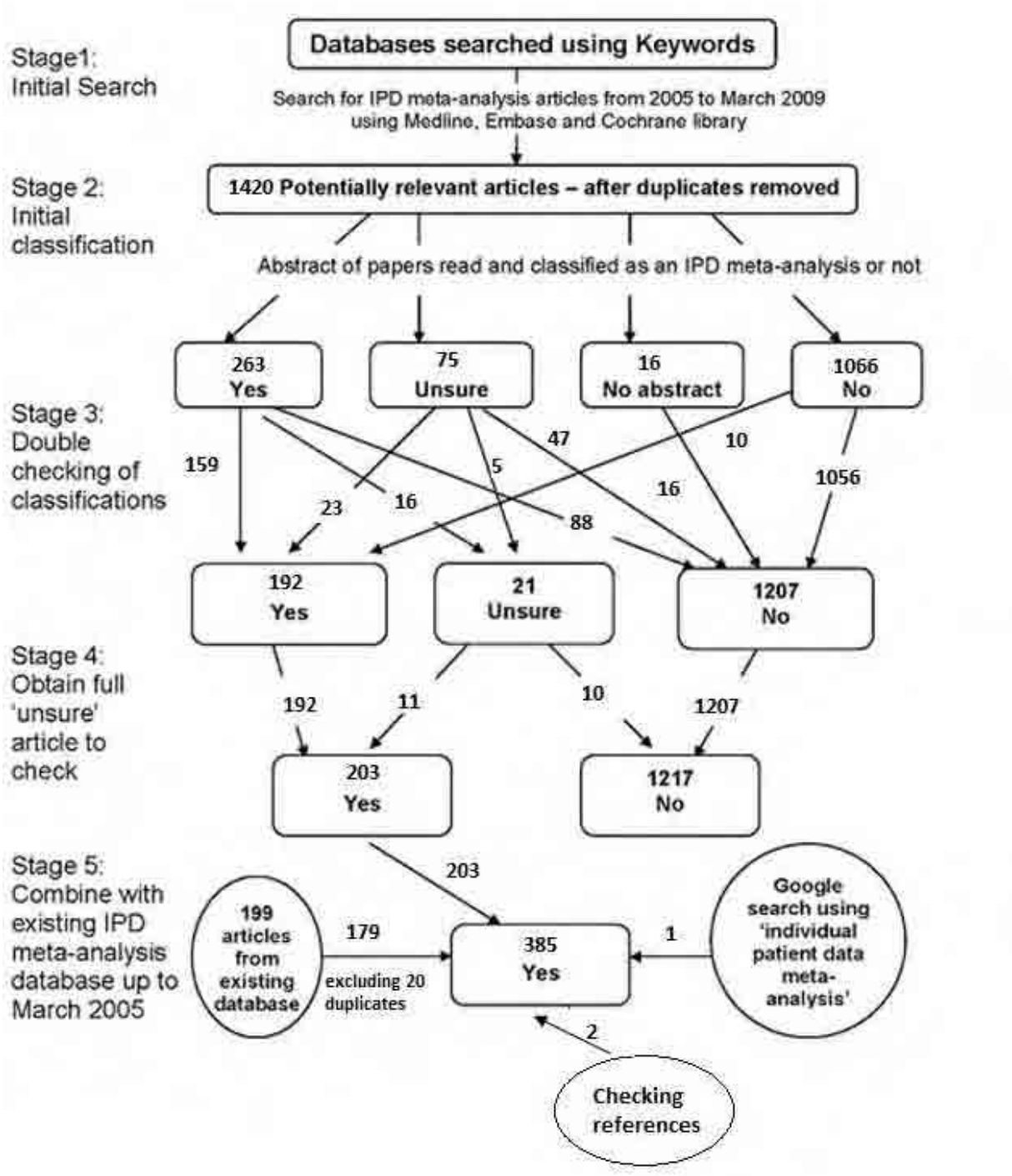
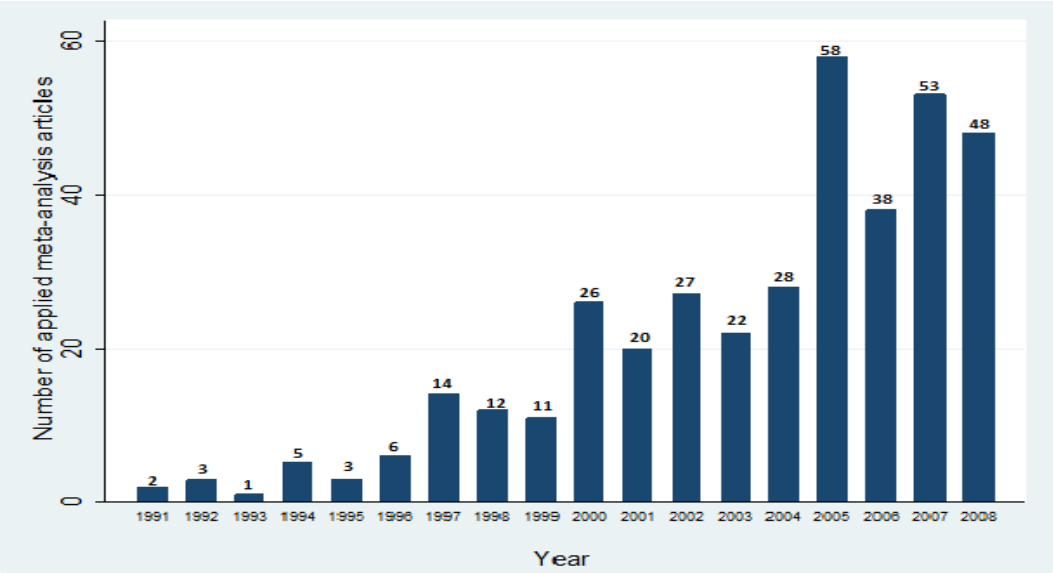


Figure 2.3: Graph showing the number of district, applied IPD meta-analysis articles published over time (6 articles were also identified in 2009, up to 5 th March when the review was conducted), as identified by my systematic review.



2.3.2 IPD Meta-analysis of Prognostic Factor Studies

The 385 IPD meta-analysis articles were classified and double checked in regards whether they focused on evaluating a PF (Figure 2.4). Ninety articles were identified that assessed an association between one or more factors and an outcome. I then distinguished between PF articles (where at baseline patients have disease or some health condition and aetiological articles (where at baseline patients are healthy).

Forty-eight of ninety articles were classified as IMPF. Figure 2.5 shows the trend of these studies over time. It can be seen from the graph that there are no studies available in 1991, 1992, 1995 and from 1997 to 1999 (also none published in 2009 up to March where the review included) . There are fluctuations of the number of PF studies over time, however, it is noticeable that these studies tend to rise from 2000 to 2008, and it appears to be an increasingly active research area.

Figure 2.4: Description and results of the search and classification of IPD meta-analysis of PF articles

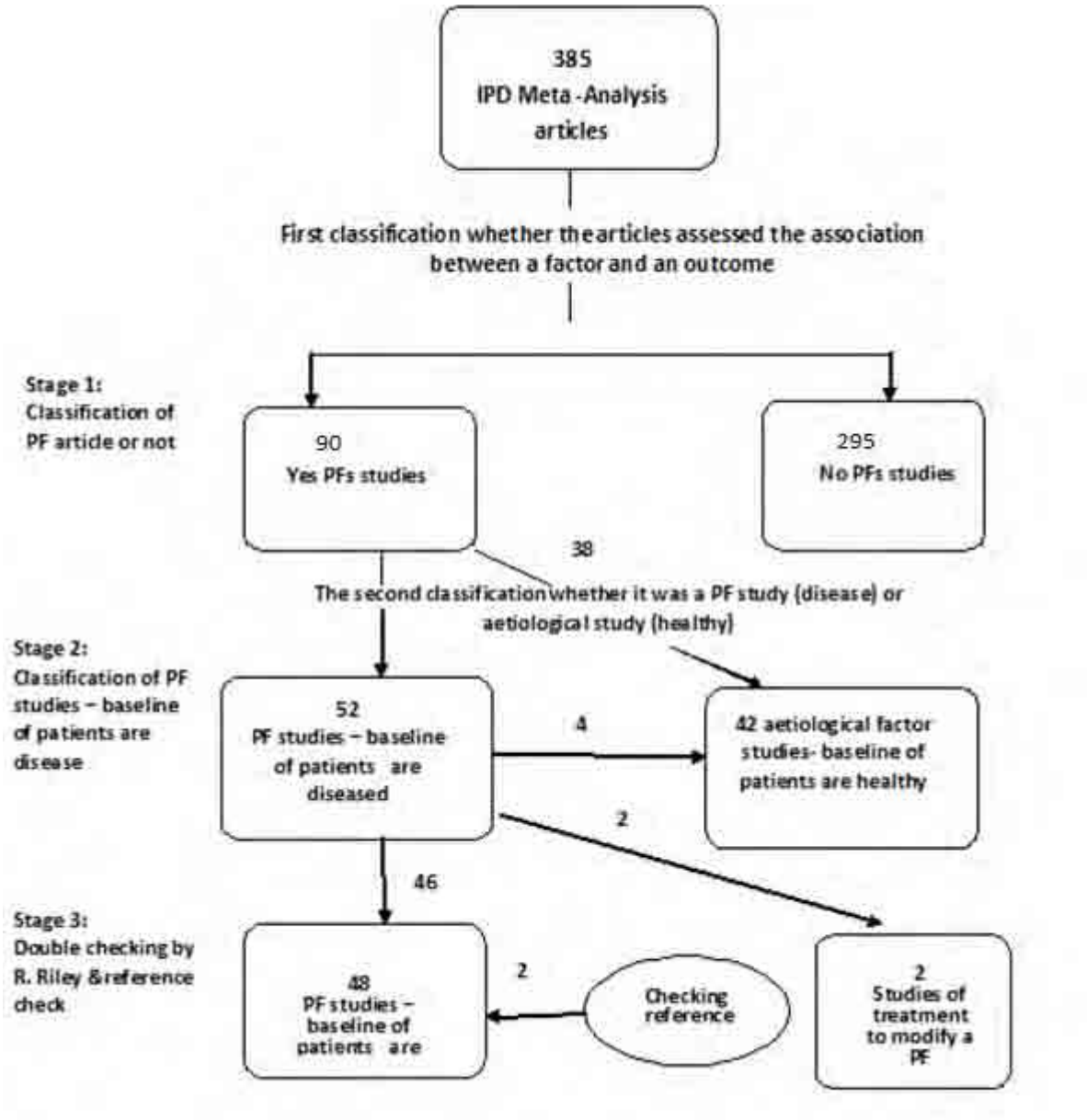
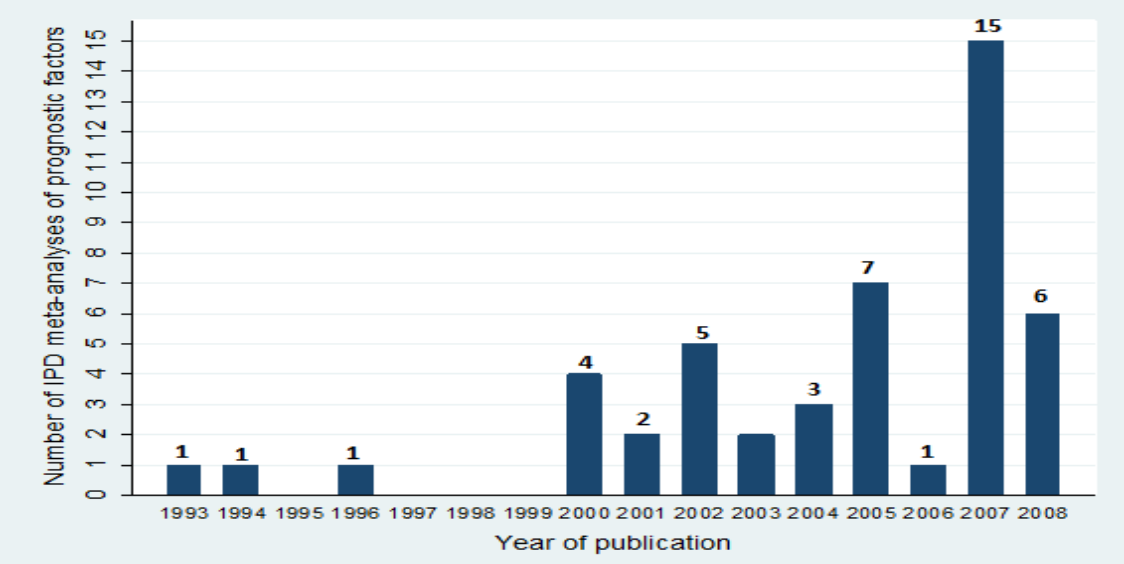


Figure 2.5: Number of published IMPF articles over time (NB no articles were identified in 2009 up to the start of March, when my review was conducted); the spike in 2007 is due to eight articles from the IMPACT collaboration being published simulaneously within the Journal of Neurotrauma



2.4 In-depth Evaluation of the 20 Most Recent IMPF Articles

In-depth evaluation was performed on a sample of 20 IMPF articles published from 2006 to 2008[72-91]. Twenty was chosen to give a manageable number of articles to review in the time available. These include:

- Six studies from IMPACT collaboration[74-79] entitled 'International Mission for Prognosis and Analysis of Clinical Trials in traumatic brain injury'. 'IMPACT' articles utilize a database of 11 studies (8 randomised trials and 3 observational studies) provided by 'IMPACT' collaborators; however I considered each IMPACT article independently and so did not, for example, assume consistent methodology in all IMPACT papers, or assume the information presented in one IMPACT paper was transferable to another IMPACT paper containing missing details.
- Nine 'literature review' articles; 'Literature review' articles sought IPD by first identifying relevant studies using a literature review (e.g. using Medline, Embase) and then contacting authors of all identified articles to ask for their IPD.

- Five articles classified under the name of 'Others' articles; the ‘Others’ articles are non-IMPACT articles that obtained IPD via existing personal articles without a literature review.

I now summarise the key findings in relation to the eight questionnaire categories defined in Figure 2.1.

2.4.1 Aims and Initiations

Table 2.2 summarises basic information about the 20 IMPF articles; including the location for the first author, the PF, the outcome of interest and whether it is a protocol driven. A common aim for the IMPF projects was to increase statistical power compared to individual studies alone, and to resolve disagreements in the field; for example, the aim for Lanterna et al. [83] was that *'emerging evidence suggests that the APOE4 allele may increase the risk of a negative outcome in patients with aneurysmal subarachnoid hemorrhage, but the results are conflicting.'* Another common objective was to identify whether a PF was still an independent PF after adjusting by some other confounding factors. For example, Trivella et al.[72] assess whether microvessel-density has prognostic value in non-small-cell lung carcinoma when adjusting for other variables including age and stage of disease.

The location of the first author for the most often the Netherlands (7) due to their leading involvement in the 'IMPACT' initiative followed by United kingdom (4). Age and gender were the most common PFs of interest in the 20's IMPF articles, and there were evaluated alongside with disease specific factors (e.g. Micro vessel-density counts in lung cancer, Palliative performance scale in heart failure). All of the 'IMPACT' studies focus on the traumatic brain injury as a disease and usually 6 month outcome in terms of Glasgow outcome scale.

Table 2. 2: Summary of the general information of the review of the random sample of IMPF studies

No.	First Author	Type of study	Year of Publication	Location of first author	Disease or condition at baseline	Main PFs of interest	Outcomes of interest	Was it stated that the IPD project was funded?	Was a protocol explicitly mentioned for the IPD project?	Was mention that ethics approval was given for the IPD project?	Was a prognostic model ('risk score') also developed in the article?
1	Butcher	IMPACT	2007	United Kingdom	TBI*	Systolic and mean arterial blood pressure	Analysis for various Glasgow Outcome Scale categories.	Yes	No	No	No
2	McHugh	IMPACT	2007	United Kingdom	TBI	Hypoxia, hypotension, and hypothermia	Glasgow Outcome Scale	Yes	No	No	No
3	Murray	IMPACT	2007	United Kingdom	TBI	Age, Glasgow Coma Scale, motor score, pupil response, computerized tomography (CT) characteristics, hypotension, hypoxia, etc.	Glasgow Outcome Scale at 6 months after injury	Yes	No	No	No
4	Maas	IMPACT	2007	Netherlands	TBI	CT classification and CT characteristics, and also other PFs as secondary analysis too	6-month outcome, assessed by the Glasgow Outcome Scale (GOS)	Yes	No	No	No
5	Mushkudiani	IMPACT	2007	Netherlands	TBI	Age , gender , race , and education	6 - month outcome, assessed by the Glasgow Outcome Scale after	Yes	No	No	No
6	Beek	IMPACT	2007	Netherlands	TBI	Laboratory parameters (e.g. glucose, sodium, haemoglobin etc)	6 months Glasgow Outcome Scale	Yes	No	No	No
7	Koopman	Literature review	2008	Netherlands	Acute otitis media	Age, sex, season, pain, fever, smoking, recurrent, etc	Middle ear effusion at 1 month	Yes	No	No	Yes
8	Thakkinstian	Literature review	2008	Thailand	Renal transplantation	Cytokinin gene polymorphism	Chronic allograft nephropathy, graft rejection, graft failure	No	No	No	No

***Traumatic Brain Injury**

Continued on next page

Continue from previous page

No.	First Author	Type of study	Year of Publication	Location of first author	Disease or condition at baseline	Main PFs of interest	Outcomes of interest	Was it stated that the IPD project was funded?	Was a protocol explicitly mentioned for the IPD project?	Was It mentioned that ethics approval was given for the IPD project?	Was a prognostic model ('risk score') also developed in the article?
9	MeRGE1	Literature review	2008	New Zealand	Acute Myocardial Infraction	Restrictive filling pattern(RFP) was the key interest; they also want to adjust for the other factors of LV systolic function, LV volumes, and Killip class (though they also adjust for age , gender and a few other variables in the full model)	Mortality	No	Yes	No	No
10	Noordzij	Literature review	2008	Boston	Monitoring for hypocalcemia after thyroidectomy	Preoperative PTH	Development of hypocalcemia	No	No	No	Yes
11	Downing	Literature review	2007	Canada	AT baseline, all patients were having palliative care (i.e. care to reduce suffering and improve quality of life); it is not disease specific but they all have severe disease of some sort	Palliative Performance Scale	length of survival	Yes	Yes	Yes	No
12	Rovers	Literature review	2007	Netherlands	Acute otities media AOM	Age, gender, smoking, recurrent AOM etc	Pain and/or fever at 3 to 7 days	Yes	No	No	Yes
13	Trivella	Literature review	2007	United Kingdom	Non-metastatic surgically treated non-small-cell lung carcinoma	Micro vessel-density counts	Overall survival	Yes	No	No	No
14	MeRGE2	Literature review	2008	New Zealand.	Heart failure	Left ventricular ejection fraction , age, gender, NYHA class and the presence of a RFP	Mortality	Yes	Yes	No	No

Continue on the next page

Continue from previous page

No.	First Author	Type of study	Year of Publication	Location of first author	Disease or condition at baseline	Main PFs of interest	Outcomes of interest	Was it stated that the IPD project was funded?	Was a protocol explicitly mentioned for the IPD project?	Was it mentioned that ethics approval was given for the IPD project?	Was a prognostic model ('risk score') also developed in the article?
15	Lanterna	Literature review	2007	Italia	Aneurysmal subarachnoid haemorrhage	APOE4 allele	The risk of a negative outcome in patients with aneurysmal subarachnoid hemorrhage and delayed ischemia, a major complication of SAH	No	No	No	No
16	Yap	Others	2007	Malaysia	Myocardial infraction	Demographic information, medical history, smoke, Concomitant medication	Death	Yes	No	No	Yes
17	Schaich	Others	2007	Germany	Acute myeloid leukaemia (AML)	Trisomy 8 (+8), either on its own or with one additional aberration in addition other PFs also of interest from clinical, cytogenetic and laboratory data	Overall survival, relapse-free survival, complete remission	Yes	No	No	Yes
18	Sylaja	Others	2007	Canada	Acute ischemic stroke	Aged >70 years, severe stroke (NIHSS >20), diabetics, history of CHF, and Hispanic ethnicity	Two outcomes of interest: SICH (symptomatic intracerebral haemorrhage) a favourable outcome (defined by a favourable function recovery score of $mRS<=1$)	No	No	No	No
19	Warkentin	Others	2006	Canada	Any medical or surgical patients being treated with heparin (either unfractionated heparin (UFH) and low molecular- weight heparin	Gender, type of heparin (UFH vs. LMWH) and type of patient (Surgical vs. medical) and the interaction between them.	Heparin-induced thrombocytopenia (HIT)	Yes	No	No	No
20	Goetz	Others	2008	Chicago	Parkinson's Disease	Patient-based factors (e.g. gender, age...etc),Study-dependent factors included medical vs. surgical interventions, likelihood of placebo assignment 50% vs. 50%, and disease severity inclusion criteria	50% improvement in total Unified Parkinson's Disease Rating Scale motor score or a decrease by 2 points on at least two UPDRSm items compared to baseline	Yes	No	No	No

End of Table (2.2)

Note that each 'IMPACT' article focused on a different PF. In the 'literature review' articles and 'Others' articles, acute otitis media and cardiovascular disease were the most common disease; whilst survival is the most common outcome.

Fifteen of the 20 articles mentioned that the project was funded. Only 3 articles[88 92 93] directly stated there was a protocol for the project, and only one mentioned they had ethics approval[85], see Table 2.2.

2.4.2 Process of obtaining IPD

Details about the process of obtaining IPD are summarized in Table 2.3. Nine of the 20 IMPF articles used a literature review to identify primary studies for which IPD was desired. Six utilised those studies already providing IPD within the IMPACT database directly[74-79]; in the 'others' articles, one utilised a set of known German trials[88]; one contacted colleagues in the field who had directed relevant trials with placebo arms[89]; one identified studies from a recent systematic review and 'from our files' [90]; and two did not state how they identified relevant studies [80 91], see Table 2.3.

In the nine articles that used 'literature reviews', just six reported the keywords used in searching, and the most common database used for searching were PubMed, Embase, the Cochrane databases, Medline and Google Scholar. Only four of the nine articles explicity stated how authors were approached for their studies (three by E-mail and one by letter); four simply said authors 'were asked', and one article did not mention anything in this regard. Only one of the nine articles reported a flowchart detailing the process of searching, classifying, and retreiving IPD studies[81].

Two of the twenty articles revealed some resource related issues for obtaining and managing the IPD. Thakkinstian et al. [81] state that '*data cleaning and checking were performed separately*

for each study', whilst more strikingly Trivella et al.[72] state that '*checking, validation and standardisation of all datasets took nearly two years'*. None of the 20 IMPF articles reported that any of the studies providing IPD held back some of their IPD.

2.4.3. Details of IPD Obtained

For the nine IMPF articles using a literature review to identify relevant IPD studies; none of these obtained IPD from all studies desired, see Figure (2.6). Six of these articles explained why the IPD was not always available: reasons included non-response to e-mail, IPD no longer available, lack of resources to participate; For example Trivella et al.[72] state that *'nine centers had no data, one had no resources to participate and 10 did not reply'*, while Rovers et al.[87]*state that 'inadequate randomization or lack of availability of information on the outcomes'*, see Table 2.3

Figure 2.6: The number of IPD studies that was requested and obtained, in each of the nine IMPF articles, using a literature review to identify relevant studies

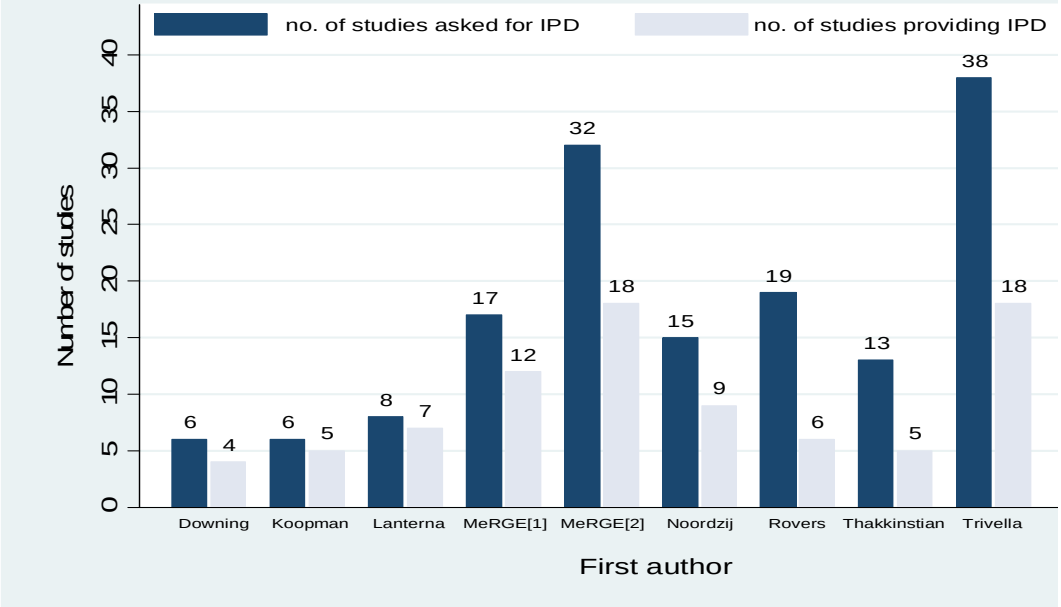


Table 2.3: Details about the process of obtaining IPD

No.	First Author	How were relevant IPD studies identified	If a literature review, was a list presented of keywords used for searching?	If a literature review, what sources were searched?	How were authors of relevant studies approached for obtaining their IPD?	What resource-related issues for obtaining IPD were mentioned?	How many studies were asked for their IPD?	How many studies provided their IPD?	What were the reasons for studies not providing IPD?	Was it stated that any of the studies providing IPD held back some of their IPD?	Was it stated that the IPD provided were different to that used by the same study in a previous publication?	Was a flowchart given showing the process of searching, classifying, retrieving and including IPD studies?	Was it stated that unpublished studies or published data were included in the IPD provided?
1	Butcher	IMPACT database	Not relevant	Not relevant	Not relevant	None mentioned	Not relevant	9	Not relevant	No	No	No	No
2	McHugh	IMPACT database	Not relevant	Not relevant	Not relevant	None mentioned	Not relevant	11	Not relevant	No	No	No	No
3	Murray	IMPACT database	Not relevant	Not relevant	Not relevant	None mentioned	Not relevant	11	Not relevant	No	No	No	No
4	Maas	IMPACT database	Not relevant	Not relevant	Not relevant	None mentioned	Not relevant	11	Not relevant	No	No	No	No
5	Mushkudiani	IMPACT database	Not relevant	Not relevant	Not relevant	None mentioned	Not relevant	11	Not relevant	No	No	No	No
6	Beek	IMPACT database	Not relevant	Not relevant	Not relevant	None mentioned	Not relevant	7	Not relevant	No	No	No	No
7	Koopman	Literature review	No	PubMed,EMBASE, Cochrane databases, and the proceedings of international symposia on recent advances in otitis media	Unclear-they were asked	None mentioned	6	5	Data were not available	No	No	No	No
8	Thakkinstian	Literature review	Yes	Medline	E-mail	Data cleaning and checking were performed separately for each centre	13	5	No reasons given	No	No	Yes	No

Continued on next page

Continue from the previous page

No	First Author	How were relevant IPD studies identified	If a literature review, was a list presented of keywords used for searching?	If a literature review, what sources were searched?	How were authors of relevant studies approached for obtaining their IPD?	What resource-related issues for obtaining IPD were mentioned?	How many studies were asked for their IPD?	How many studies provided their IPD?	What were the reasons for studies not providing IPD?	Was it stated that any of the studies providing IPD held back some of their IPD?	Was it stated that the IPD provided were different to that used by the same study in a previous publication?	Was a flowchart given showing the process of searching, classifying, Retrieving and including IPD studies?	Was it stated that unpublished studies or published data were included in the IPD provided?
9	MeRGE2	Literature review	Yes	Biological Abstracts, Clinical Evidence, Current Contents, Embase, Medline, Medline In-progress and PubMed.	Unclear-they were asked	None mentioned	17	12	Some of them inaccessible	No	No	No	No
10	Yap	Others	No	Not stated	Not stated	Not mentioned	4	4	Not relevant	No	No	No	No
11	Rovers	Literature review	no	PubMed, Embase,the proceedings of International Symposia on Recent Advances in Otitis Media, and the Cochrane	Unclear-they were asked	None mentioned	19	6	Inadequate randomization or lack of availability of information on the outcomes	No	No	No	No
12	Trivella	Literature review	Yes	Medline	Letter	High cost and time consuming (checking, validation, and standardization of all datasets took nearly 2 years)	38	18	Nine centres had no relevant data;1 stated they had no resources to participate; 10 did not reply	No	No	No	Yes
13	MeRGE1	Literature review	Yes	Biological Abstracts, Clinical Evidence, Current Contents,Embase, Medline,Medline In-progress and PubMed,	Not mentioned	Not mentioned	32	18	Not-accessible	No	No	No	Yes
14	Noordzij	Literature review	Yes	Pubmed	E-mail	None mentioned	15	9	No reason given	No	No	No	No

Continued on next page

Continued from previous page

NO.	First Author	How were relevant IPD studies identified	If a literature review, was a list presented of keywords used for searching?	If a literature review, what sources were searched?	How were authors of relevant studies approached for obtaining their IPD?	What resource-related issues for obtaining IPD were mentioned?	How many studies were asked for their IPD?	How many studies provided their IPD?	What were the reasons for studies not providing IPD?	Was it stated that any of the studies providing IPD held back some of their IPD?	Was it stated that the IPD provided were different to that used by the same study in a previous publication?	Was a flowchart given showing the process of searching, classifying, Retrieving and including IPD studies?	Was it stated that unpublished studies or published data were included in the IPD
15	Lanterna	Literature review	Yes	M EDLINE, EMBASE,the Cochrane Library,CINHAL, and LILACS	Unclear-'they were asked'	None mentioned	8	7	No reason given	No	No	No	No
16	Downing	Literature review	No	Medline database, Web of science and Google scholar	E-mail	Not mentioned	6	3*	Yes-2 'did not reply' and 1 did not have access to the original data anymore	No	No	No	No
17	Warkentin	'Others'-Studies already identified from recent systematic reviews and from our files	Not relevant	Not relevant	Unclear	None mentioned	8	7	They omitted one study because it has not gender information which considered the main factor in this study	No	No	No	No
18	Sylaja	Others	-	-	Not mentioned	None mentioned	Unclear	4	None mentioned	No	No	No	No
19	Schaich	'Others'- Known German trials involving acute myeloid leukaemia patients between April 1993 and December 2002	Not relevant	Not relevant	Not mentioned	None mentioned	8	8	Not relevant	No	No	No	No
20	Goetz	'Others'-contacting colleagues who had directed relevant clinical trials with placebo arms	Not relevant	Not relevant	Unclear, they were 'contacted'	Not mentioned	Unclear	11	Not mentioned	No	No	No	No

- Four researchers response to provide their IPD, but only three provide the IPD studies

End of Table 2.3

The percentage of studies providing IPD ranged from 32% to 88%, and five of the nine articles obtained IPD from 60% or less of the requested studies. The shortfall appeared larger in those IMPF articles requesting IPD from 10 or more studies see Figure 2.6.

In the 'others' five IMPF articles, one article obtained IPD from all studies desired[88], and only one of these five articles reported the reason for not providing all IPD studies[90].

Table 2.3 gives details about included IPD studies for the 20 IMPF articles. All of the 20 articles provided the number of patients included in their available IPD (Figure 2.7); this ranged from 131 to 8721, indeed, IMPACT studies have the highest number of patients included in their availability IPD, this range between 5672 and 8721. Sixteen articles also reported the number of patients separately for each included IPD study, but only 4 articles reported the number of outcome events separately for each IPD study.

Figure 2.7: The number of patients included in IPD studies for the 20 IMPF articles

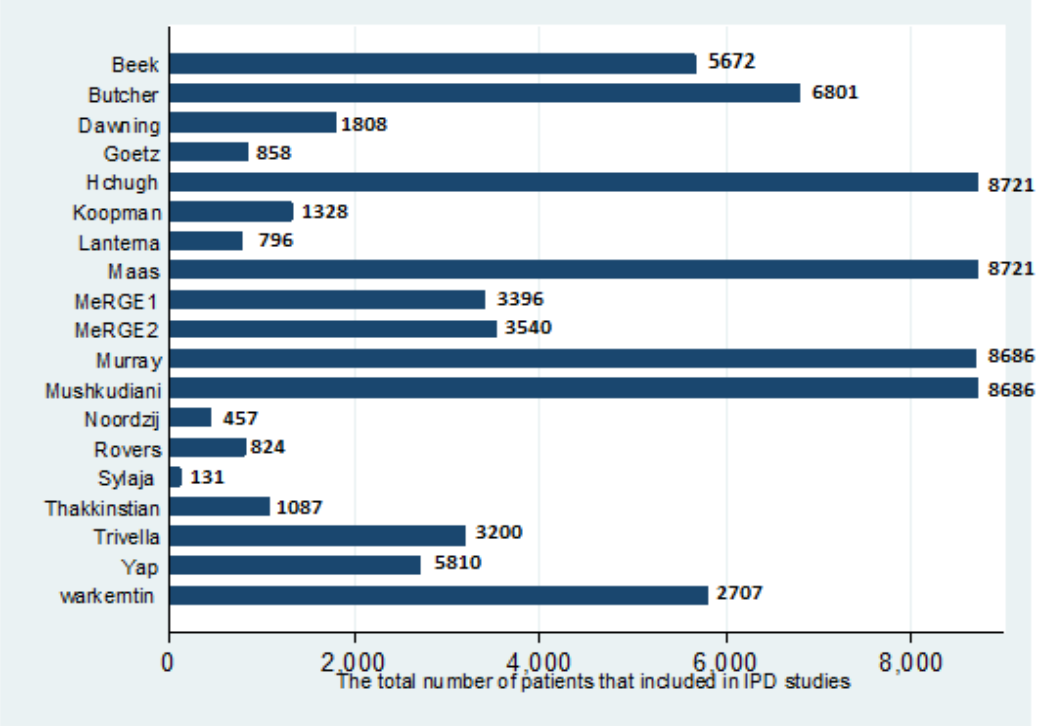


Table 2. 4: Details about included IPD studies for the 20 IPFM articles

Study ID	First Author	Type of the study	How many studies provided their IPD?	What types of studies were they (how many of each)?	Across all the IPD studies, How many patients were included in the IPD provided?	Was the number of patients reported for each IPD study separately ?	Was the number of events reported for each IPD study separately?	What problems of missing data were there in the IPD database?	Was a quality assessment done for each included study providing IPD (e.g. to check the quality of the study design, the process used to follow-up and collate data, etc)
1	Butcher	IMPACT	9	8 RCTs, 1 observational Study	6801	Yes	No	Missing data and missing outcomes	No
2	McHugh	IMPACT	11	8 RCTs and 3 observational studies	8721	Yes	No	Missing data and missing outcomes	No
3	Murray	IMPACT	11	There are 3 observational studies and 8 RCTs	8686	No	No	Missing values	No
4	Maas	IMPACT	11	11- but it doesn't state in this paper whether they are RCTs or what (in a related paper no.14 it tell us they are 8 RCTs and 3 observational studies, but it never mentioned that here)	8721	Yes	No	Missing data and missing variables	No
5	Mushkudiani	IMPACT	11	Eight therapeutic Phase III randomized clinical trials and three surveys in moderate or severe TBI	8721	Yes	No	Missing variables and missing outcomes	No
6	Beek	IMPACT	7	6 Randomized trials and 1 observed	5672	Yes	No	Missing values, missing variables and missing outcomes.	No
7	Koopman	Literature review	5	All randomized trials	1328	Yes	No	Missing variables, missing values and missing outcomes	No
8	Thakkinstian	Literature review	5	'Human population-based association studies'	1087	Yes	No	Missing variables and missing outcomes	The Hardy-Weinberg equilibrium (HWE) was assessed in each study providing IPD, only studies that observed HWE were included in the analysis
9	MeRGE 2	Literature review	12	12 prospective studies including 5 clinical trials	3396	No	No	Missing values	No

Continue on next page

Continue from previous page

Study ID	First Author	Type of the study	How many studies provided their IPD?	What types of studies were they (how many of each)?	Across all the IPD studies, How many patients were included in the IPD provided?	Was the number of patients reported for each IPD study separately?	Was the number of events reported for each IPD study separately?	What problems of missing data were there in the IPD database?	Was a quality assessment done for each included study providing IPD (e.g. to check the quality of the study design, the process used to follow-up and collate data, etc)
10	Yap	Others	4	All of the Randomized trials	2707	Yes	Yes	Missing data and missing variables	No
11	Rovers	Literature review	6	All of the randomized trials	824	No	No	Missing data, missing values and missing variables	No
12	Trivella	Literature review	18	Not relevant - data provided by collaborating centres rather than specific studies	3200	Yes	No	Missing variables	No
13	MeRGE1	Literature review	18	Of these 18 studies, it is not clear what type there were, other than that they were 'prospective'	3540	No	No	Missing data	No
14	Noordzij	Literature review	9	Observational studies	457	Yes	No	Missing variables	No
15	Lanterna	Literature review	7	Not clear some prospective and some retrospective studies.	796	Yes	Yes	Missing data and missing outcomes	No
16	Downing	Literature review	4	Cohort studies	1808	Yes	No	Missing variables, but they did focus on the available variables.	Yes, using quality assessment criteria proposed by McKibbon in 1998, and Altman in 2001
17	Sylaja	Others	4	All cohort studies	1966	Yes	No	Missing outcome	No
18	Schaich	Others	8	All are randomized trials	131	Yes	No	Missing values	No
19	Warkentin	Others	7	7 ; 2 randomized trials and 5 prospective cohort studies	5810	Yes	Yes	Missing data and missing variables	No
20	Goetz	Others	11	All of the randomized trials	858	Yes	Yes	Missing data	No

End of Table 2.4

Missing data was a major problem within the IPD obtained both at the patient-level and at the study-level. All of the 20 IMPF articles reported at least one of missing data problem [Table 2.4]; in particular, missing values of PFs and adjustment factors for some patients within a study (e.g. Murray et al.[74]); missing outcome data for some patients within a study (e.g. Sylaja et al.[91]) ; and some PFs or outcome missing completely in some studies (e.g.Trivella et al.[72], Lanterna et al.[83]); only 9 of the 20 IMPF articles reported use of one of the statistical models (e.g. imputation analysis) to limit the missing data problem.

2.4.4 Type and quality of IPD studies

Over all of the 20 IMPF articles, 3 articles did not report directly what type of IPD studies was included (e.g. randomised trials, observational studies etc.). For example, Lanterna et al.[83] stated that '*There were some prospective studies and retrospective studies*' but the author did not go into specific details. Seventeen articles reported the type of IPD studies included; six of them were IMPACT articles that utilised IPD from 8 randomised trials and 3 observational studies. Five articles utilised IPD from randomised trial only (for example Yap et al.[80] utilised IPD from just the placebo arm of randomised trials); two articles utilised IPD from randomised and prospective studies; two articles utilised IPD from cohort studies; one article utilised IPD from observational studies[86] and one article utilised IPD from human population- based association studies[81], see Table 2.4.

Only two of the twenty articles mentioned that some of their IPD studies were unpublished [72] [74] , and only two studies reported a quality assessment for each included study providing IPD; for example, Thakkinstian et al. [81] assessed the Handy-Weinberg Equilibrium (HWE) in each study providing IPD, and only the studies that observed HWE were ultimately included in the project. Downing et al.[85] used quality assessment criteria proposed by Mckibbon[66] in 1998 and Altman[72] in 2001.

2.4.5 Statistical Methods Used

All of the 20 articles provided a statistical analysis description in their methods section, see Table 2.5.

Statistical Models for the Patient-level Data

Logistic regression and Cox proportional hazard model were the most common statistical models used to analyze the patient level data; in particular 6 of them used Cox proportional hazard model, 10 articles used either logistic or proportion odds regression and one article used both Cox regression and logistic regression[89], see Table 2.5. Two of the twenty articles did not use any model, for example in sylaja et al.[91] the authors did direct calculations of proportions in each group defined by the PF. Only three of the 20 articles reported checking model assumptions, all in regard the proportional hazards assumption in Cox regression. In two of these three articles[85 92], it was explicitly stated that model assumptions were checked separately in each study included in the meta-analysis.

Meta-analysis Framework

For meta-analysis, seven articles used a two-step approach where the IPD was firstly analysed in each study separately, and then in the second step the summary data obtained (e.g. hazard ratios, odds ratios) were synthesised using a traditional model for meta-analysis of aggregate data (see chapter 1.7). Ten articles used a one-step approach, where the IPD across all studies were analysed together simultaneously; one article considered both one-step and two step methods[89]. In the two remaining articles the choice of one-step or two-step method was not clear.

In the one-step method, there are two approaches for estimating the effect size of the PF; the first approach accounts for the clustering of patients within studies by either including a

dummy variable for the study or by using multilevel modelling[66 94]; the second approach does not account for the clustering across studies by assuming all of the IPD studies came from one study. In contrast, the two-step method accounts naturally for clustering of patients within studies (see chapter 4). In five of the ten articles that used a one-step method, the authors explicitly reported that they accounted for the clustering of patients within studies (Table 2.5). The other 5 may have analysed the IPD as it came from one study, although this was not stated explicitly.

Assessing and Accounting for Heterogeneity

Details of methods for assessing and accounting for heterogenity in PF effect are summarized in Table 2.5. Between-study heterogeneity in prognostic effects (e.g. hazard ratios, odds ratios) was examined in 14 of the 20 articles, typically using the I^2 statistic [95] or the Q-statistic (chi-square test for heterogeneity)[96]. Of the 20 articles, 8 accounted for heterogeneity by including random-effects; seven did not account for heterogeneity and justified why, for example as it was negligible (e.g. Koopman et al.[82] noted small values of $I^2 < 25\%$) or the chi-square test was non-significant (e.g. Thakkinstian et al.[81] noted the test gave $p > 0.1$), see Table 2.5; five articles did not account for heterogeneity in their meta-analysis and did not explain why; and in the remaining article heterogeneity was examined but it was unclear whether it was accounted for in the meta-analysis . Three articles [72 83 90] examined potential causes of between-study heterogeneity; for example, Trivella et al.[72] perform subgroup analyses according to the method of measuring microvessel density; Lanterna et al.[83] used random-effects meta-regression; and Warkentin et al.[90] looked at excluding a 'before and after' prospective cohort study that they believe is causing the heterogeneity.

Table 2. 5: Details about the summary of the statistical methods over the 20 IMPF articles

Study ID	First Author	Was a statistical analysis description provided in the Methods section?	Did they state that their meta-analysis method accounted for clustering of patients within studies?	What general structure was used for the meta-analysis (i.e. 'one-step' ignoring clustering; 'one-step' accounting for clustering; 'two-step')?	What statistical models were used at the patient-level in the meta-analysis?	What model assumptions were checked in relation to these models? [Note if they were done in each study separately. Or at the synthesis level]?	Was between-study heterogeneity in the PF effect assessed? If yes, how?	Was between-study heterogeneity in the prognostic effect accounted for in the meta-analysis? If yes - how? If no, why not?	Were causes of between-study heterogeneity examined, and if so, how?	Did the meta-analysis consider the quality of available IPD studies (e.g. sensitivity analysis excluding low quality studies)?
1	Butcher	Yes	Yes	'Two -step'	Proportional odds and logistic regression models	None checked	No assessment of heterogeneity	'No because they state prognostic effects 'demonstrated considerable consistency from study to study'	No	Not relevant
2	McHugh	Yes	Yes	'Two-step'	Binary and proportional odds regression models	None checked	Yes but they did not say how	Yes random effect pooled	No-presumably as consistent effect across studies	Not relevant
3	Murray	Yes	Yes	'Two-step'	Multiple logistic regression models,	None checked	No assessment of heterogeneity	Yes, using a random-effects model	Not relevant as there is no heterogeneity	Not relevant as no quality criteria applied
4	Maas	Yes	Yes	'Two-step'	Binary logistic and proportional odds regression model	None checked	Yes, by unnamed test of heterogeneity	Yes, using a random-effects model	No	Not relevant as not quality criteria applied
5	Mushkudiani	Yes	Yes	'Not clear'	Logistic regression and proportional odds regression	None checked	No assessment of heterogeneity	Yes, using a random-effects model	No	Not relevant as no quality criteria applied
6	Beek	Yes	No	'Not clear'	Logistic and proportion odds regression	None checked	Yes,they present a p-value from an unnamed test for heterogeneity	'Not clear'	No	Not relevant as no quality criteria applied
7	Koopman	Yes	No	'One-step'	Logistic regression	None checked	Yes, by using I^2	No, because $I^2 < 25\%$	Not relevant, as heterogeneity was considered small	No
8	Thakkinstian	Yes	Yes	'One-step'	Logistic regression	None checked	Yes, by using a chi-square test	No, because p-value from chi-square test > 0.1	Not relevant, as heterogeneity was considered small	Yes-studies failing the Hardy-Weinberg equilibrium were excluded
9	MeRGE 2	Yes	Yes	'One-step'	Cox PH model	None checked	Yes, by using I^2 and chi-square test	No, because $I^2 = 0.042$ and p-value from chi-square test is 0.06, which they deem to indicate little heterogeneity*	Not relevant as heterogeneity considered small'	Not relevant as studies were not classed as low or high quality

Study ID	First Author	Was a statistical analysis description provided in the Methods section?	Did they state that their meta-analysis method accounted for clustering of patients within studies	What general structure was used for the meta-analysis (i.e. 'one-step' ignoring clustering; 'one-step' accounting for clustering; 'two-step')?	What statistical models were used at the patient-level in the meta-analysis?	What model assumptions were checked in relation to these models? [Note if they were done in each study separately. Or at the synthesis level]?	Was between-study heterogeneity in the PF effect assessed? If yes, how?	Was between-study heterogeneity in the prognostic effect accounted for in the meta-analysis? If yes - how? If no, why not?	Were causes of between-study heterogeneity examined, and if so, how?	Did the meta-analysis consider the quality of available IPD studies (e.g. sensitivity analysis excluding low quality studies)?
10	Yap	Yes	Yes	'One-step'	Multiple Cox PH model	Proportional hazards assumptions checked; not clear if this was done for each study though	Yes, by looking at the interaction between the prognostic effect and study in the model	Yes they looked at interaction between effect and studies, but it was usually not important	Not relevant, as heterogeneity was considered small	Not relevant as studies were not classed as low or high quality
11	Rovers	Yes	No	'One-step'	Logistic regression	None checked	Yes , by using I^2	No, as they identify low heterogeneity	Not relevant as low heterogeneity	Not relevant as no quality criteria applied
12	Trivella	Yes	Yes	'Two-step'	Cox regression	None checked	Yes, by using I^2	Yes, random-effects inverse-variance meta analysis.	They split the analysis by the two different methods to measure the main variable	Not relevant as no quality criteria applied
13	MeRGE1	Yes	No	'One-step'	Cox PH model	The assumption of proportionality of hazards for RFP was assessed using Schoenfeld residuals and considered acceptable	Yes by using chi-square test and I^2	No, because $I^2 = 0$	Not relevant as there is no heterogeneity	Not relevant as no quality criteria applied
14	Noordzij	Yes	No	'One-step'	Linear regression (analysis of variance)	None checked	No assessment of heterogeneity	No	Not relevant as there is no heterogeneity	Not relevant as no quality criteria applied
15	Lanterna	Yes	Yes	'Two-step'	No models used; just direct calculation of tests	Not relevant	Yes, by Q statistic and I^2 tests	Yes $(when I^2 > 50\%$ or p-value for Q statistic $< 0.1)$, using random-effects model	Random-effect meta-regression	Not relevant as no quality criteria applied

study ID	First Author	Was a statistical analysis description provided in the Methods section?	Did they state that their meta-analysis method accounted for clustering of patients within studies?	What general structure was used for the meta-analysis (i.e. 'one-step' ignoring clustering; 'one-step' accounting for clustering; 'two-step')?	What statistical models were used at the patient-level in the meta-analysis?	What model assumptions were checked in relation to these models? [Note if they were done in each study separately. Or at the synthesis level]?	Was between-study heterogeneity in the PF effect assessed? If yes, how?	Was between-study heterogeneity in the prognostic effect accounted for in the meta-analysis? If yes - how? If no, why not?	Were causes of between-study heterogeneity examined, and if so, how?	Did the meta-analysis consider the quality of available IPD studies (e.g. sensitivity analysis excluding low quality studies)?
15	Lanterna	Yes	Yes	'Two-step'	No models used; just direct calculation of tests	Not relevant	Yes, by Q statistic and I^2 tests	Yes ($when I^2 > 50\%$ or p-value for Q statistic < 0.1), using random-effects model	Random-effect meta-regression	Not relevant as no quality criteria applied
16	Downing	Yes	Yes	'One-step'	Cox PH model	Yes, by using martingale, schoenfeld, and deviance statistic within each study	Yes, they visually assess whether between-study heterogeneity exists	No, because no evidence of heterogeneity	Not relevant	Not relevant as no quality criteria applied
17	Sylaja	Yes	No	'One-step'	No models used; just direct calculations of proportions in each group defined by the prognostic factor	Not relevant	No assessment of heterogeneity	No	Not relevant because heterogeneity was not assessed	Not relevant as no quality criteria applied
18	Schaich	Yes	Yes	'One-step'	Cox regression and logistic regression	None checked	No assessment of heterogeneity	No	Not relevant as heterogeneity not assessed'	Not relevant as no quality criteria applied
19	Warkentin	Yes	Yes	'Two-step'	Random-effects logistic model	None checked	Breslow- day statistic	Yes, using random-effects model if p< 0.05 for the test of heterogeneity	They look at excluding a 'before-and-after' prospective cohort study that they believe is causing it	Not relevant as no quality criteria applied
20	Goetz	Yes	Yes	'One-step' in some analysis, 'two-step' in other analysis	Generalized estimating equations (GEE) for binary repeated measures	None checked	No assessment of heterogeneity	No	Not relevant as heterogeneity does not assessed	Not relevant, as no quality assessment done

End of table 2.5

Analysis of Continuous Factors

Details about the examination of continuous PFs for the 20 IMPF articles are summarized in Table 2.6. Nineteen of the 20 articles investigated one or more factors measured on a continuous scale. Of these, eight categorized the continuous factors for the analysis; five analysed the continuous factors on a continuous scale; and six used a continuous scale in some analyses (or for some factors) and used categorise in other analyses (or for other factors).

Of the eleven articles that used continuous scales, seven of them transformed the scale for the PF before analysis, because of different methods of measurements across studies which limit using the data for analysis, and skewness in the PF which hinder the linearity assumptions. For example Beek et al.[76] stated that '*for continuous PFs with a linear relation to outcome, the odds ratios were scaled so that they correspond to changing from the 25 th percentile of that PF to the 75 th percentile. This was done to allow a direct comparison of different PFs, which are recorded in different units or on different scales'*; and Yap et al.[80] stated that '*variables with skewed distribution were log-transformed',* see Table 2.6. Six of these eleven articles considered non-linear trend in their analysis; the most common non-linear model was spline model used in four articles, one article used a polynomial term and one article used a quadratic term (Table 2.6).

.

Table 2. 6: Details about treatment the continuous variables over the 20 IMPF articles

Study ID	First Author	Were continuous PFs assessed?	If yes, were they examined on a continuous scale or on a categorised scale (e.g. dichotomized)?	If on a continuous scale, was a transformed scale used and if so why?	If on a continuous scale, were non linear trends considered and, if so, how?	If on a categorized scale, were categories (i.e. cut-points) justified? If yes, give details on the justification?	Was the independent value of PFs assessed; i.e. were multivariable analyses done that examine the PF of interest adjusting for other know prognostic factors?	If yes multivariable analysis were used, were set criteria used to determine whether a PF remained important even after adjusting for other variables? If yes, what were they?	In multivariable analysis, were full results shown for the PFs of interest regardless of their statistical significance?	Was the interaction between two or more PF considered as a prognostic variable?
1	Butcher	yes	Categorized and continuous scales both used	No	Yes, using spline functions	Yes, as indicated by the spline function	Yes	No criteria used; all significant and non-significant factors remained in the model	Yes	No
2	McHugh	yes	Categorized scale	Not relevant	Not relevant	Unclear, but authors mention they were restricted by the definitions used in the original studies	Yes	No criteria mentioned	Not relevant, as all variables are significant	Yes
3	Murray	Yes	Categorized scale for some, continuous scale for others	Yes; for continuous PFs with a linear relation to outcome, the odds ratios were scaled so that they correspond to changing from the 25 th percentile of that PF to the 75 th percentile. This was done to allow a direct comparison of different PFs, which are recorded in different units or on different scales.	No, as they note relationship was linear	Yes, to allow a direct comparison of different PFs, which are recorded in different units or on different scales.	Yes	No; $p < 0.01$ used for significance, but all variables remained in the model	Yes	No
4	Maas	Yes	Categorized scale and continuous scale	No	Yes, using spline functions	No	Yes	No criteria used; all significant and non-significant factors remained in the model	Not relevant, as all variables are significant	No
5	Mushkudiani	Yes	Continuous scale	No	Yes, using spline functions	Not relevant	Yes	No criteria used; all significant and non-significant factors remained in the model	Yes	Yes
6	Beek	Yes	Continuous scale	Yes; for continuous PFs with a linear relation to outcome, the odds ratios were scaled so that they correspond to changing from the 25 th percentile of that PF to the 75 th percentile. This was done to allow a direct comparison of different PFs, which are recorded in different units or on different scales.	Yes, using spline functions	Not relevant	Yes	No criteria used; all significant and non-significant factors remained in the model	Yes	No

Continue on next page

Continue from the previous page

Study ID	First Author	Were continuous PFs assessed?	If yes, were they examined on a continuous scale or on a categorised scale (e.g. dichotomized)	If on a continuous scale, was a transformed scale used and if so why?	If on a continuous scale, were non linear trends considered and, if so, how?	If on a categorized scale, were categories (i.e. cut-points) justified? If yes, give details on the justification	Was the independent value of PFs assessed; i.e. were multivariable analysis done that examine the PF of interest adjusting for other know prognostic factors	If yes multivariable analysis were used, were set criteria used to determine whether a PF remained important even after adjusting for other variables? If yes, what were they?	In multivariable analysis, were full results shown for the PFs of interest regardless of their statistical significance?	Was the interaction between two or more PF considered as a prognostic variable?
7	Koopman	Yes	Categorized scale	Not relevant	Not relevant	No	Yes	P-value <= 0.05 meant the PF was kept in the model	No	No
8	Thakkinstian	Yes	Categorized scale	Not relevant	Not relevant	No	Yes	P-value <= 0.05 mean the PF was kept in the model	Yes	No
9	MeRGE 2	Yes	Categorized scale for one continuous variable, and continuous scale for another continuous variable (age)	Age treated per 10 years increase	No	Yes, cut-points chosen to make results 'clinically applicable'	Yes	P-value <0.1 meant that the PF was included	No	Yes
10	Yap	Yes	Continuous scale	Yes, variables with skewed distributions were log-transformed.	Yes, by adding polynomial terms	Not relevant	Yes	P-value < 0.05 meant the PF was kept in the model	No	No
11	Rovers	Yes	Categorized scale	Not relevant	Not relevant	Yes, based on previous clinical information - note that that the categorised scale was imposed on them as some studies only provided prognostic factors dichotomised and not on their original scale	Yes	P-value <= 0.05 meant the prognostic factor was kept in the model	Yes for some but not all	Yes
12	Trivella	Yes	Categorised scale and continuous scale	Sometimes the prognostic factor effect was expressed per 10 unit increase	Yes, using quadratic terms	Yes, by noting that this was to enable a comparison with the analysis leaving the factor in the continuous scale and noting the choice of -point is different across previous studies	Yes	Not criteria used	Yes	No
13	MeRGE1	Yes	Categorized scale	Not relevant	Not relevant	Yes, so that 'relevant' hazard ratios can be calculated	Yes	No criteria used; all significant and non-significant factors remained in the model	Yes	Yes

Continue on next page

Study ID	First Author	Were continuous PFs assessed?	If yes, were they examined on a continuous scale or on a categorised scale (e.g. dichotomized)?	If on a continuous scale, was a transformed scale used and if so why?	If on a continuous scale, were non linear trends considered and, if so, how?	If on a categorized scale, were categories (i.e. cut-points) justified? If yes, give details on the justification?	Was the independent value of PFs assessed; i.e. were multivariable analyses done that examine the PF of interest adjusting for other know prognostic factors?	If yes multivariable analysis were used, were set criteria used to determine whether a PF remained important even after adjusting for other variables? If yes, what were they?	In multivariable analysis, were full results shown for the PFs of interest regardless of their statistical significance?	Was the interaction between two or more PF considered as a prognostic variable?
14	Noordzij	Yes	Continuous scale	Yes, they changed to a percentage scale when modelling the percentage change from baseline, to overcome different methods of measurement across studies	No	Not relevant	No	Not relevant	Not relevant	Not relevant as only one factor considered
15	Lanterna	Yes	Categorized scale	Not relevant	Not relevant	No	No	Not relevant as no multivariable model used	Not relevant	No
16	Downing	Yes	Categorized scale	Not relevant	Not relevant	Yes, categories chosen based on those determined in a previous report	Yes	No criteria used; all significant and non-significant factors remained in the model	Yes	Yes
17	Sylaja	Yes	Categorized scale	Not relevant	Not relevant	Yes, but contradictory argument: they used a cut-off of 70 years due to other studies showing advancing age increases risk in a continuous fashion	No	Not relevant, as no multi-variable model used	Not relevant	No
18	Schaich	Yes	Categorized scale and continuous scale	No	No	Yes, as the cut-points gave a significant result in the analysis hierarchical cluster analysis	Yes	Non-significant' variables were excluded	No	No
19	Warkentin	No	Not relevant	Not relevant	Not relevant	Not relevant	Yes	No criteria used; all significant and non-significant factors remained in the model	Yes	Yes
20	Goetz	Yes	Continuous scale	Yes, e.g. age treated per 10 years increase; some other variables also transformed similarly	No	Not relevant	No	Not relevant as no multi-variable model fit	Not relevant	No

End of Table 2.6

Of the 14 articles that used categorised scales, nine articles used justification for their cut-off points; justifications were (i) to make results clinically applicable[82]; (ii) to allow a direct comparison of different PFs, which are recorded in different units or on different scales[74], (iii) based on previous clinical information[85 87]. For example, Rovers et al.[87] stated that '*some predictor and outcome variables (e.g. fever and pain) might have been more informative if analyzed on a continuous scale. Some trials did measure these items on a continuous scale but, because others did not, we needed to recode these items as dichotomous variables'*. In contrast, four of these 14 articles did not justify their choice of categories, and one article was not clear.

Assessment of Independent Prognostic Value

Sixteen of the 20 articles reported a multivariable analysis to examine the independent prognostic value of one or more factors after adjusting for others (Table 2.6). Seven out of these 16 articles defined the criteria that used to judge whether the primary PF was still an independent PF even after adjusting by other confounding factors. In particular the P- value was used to decide whether the PF and the confounding factors are significant or not; for example in Koopman et al.[82] a criteria of $p < 0.05$ was used for statistical significance in the multivariable model and thus evidence of independent prognostic value. Whilst 9 of the 16 articles did not define the criteria to decide whether the PF still independent or significant even after adding other confounding factors; for instance, in Mass et al.[77] all significant and non-significant factors remained in the model. Nine of the sixteen articles reported the results in full (whether the PF is still significant factor after adjusting by other confounding factors, i.e. by using P-value); in two articles of these 16 articles, all of the factors are significant (Table 2.6.); and five articles did not report the results in full (i.e. adjusted effect estimate with uncertainty or p-value) for those factors deemed not to have independent prognostic value.

Assessment of Interactions between Prognostic Factors

Of the sixteen articles that reported a multivariable analysis to examine the independent prognostic value of one or more factors after adjusting for others; seven of these articles considered a possible interaction between two or more factors in their multivariable model (Table 2.6). For example, McHugh et al. [79] consider whether the interaction between hypoxia and hypotension is prognostic for 6-month outcome in patients with traumatic brain injury.

2.4.6 Assessment of Publication and Availability Biases

Details about if and how IMPF articles asses publication and availability bias, and whether including non-IPD studies in the meta-analysis significantly changed conclusion about a PF, are shown in Table 2.7.

Two of the twenty articles formal assessed whether publication bias or small-study effects (i.e. the tendency for small studies in the meta-analysis to give more favourable prognostic effects than the larger studies) may be affecting their meta-analysis. For example; Lanterna et al.[83] used funnel plot for asymmetry alongside a regression asymmetry test, the funnel plot showed no asymmetry, whilst in MeRGE1 collaboration[92] the authors used a funnel plot and the authors stated *that 'The present result may be subject to publication bias. However, inspection of the funnel plot of the individual hazard ratios for each included study failed to identify important heterogeneity (*$I^2 = 0.042$ *,* $P = 0.06$*)'*

Availability bias

Ignoring the six 'IMPACT' studies (which utilised an existing IPD database), 10 of the remaining 14 articles did not obtain IPD from all desired studies (Table 2.7), although two articles did not mention whether all desired IPD studies were obtained or not[89 91].

Of the 10 articles that stated they did not obtain all the IPD desired, four studies gave the number of patients in the non-IPD studies, and only one article provided the number of events in the non-IPD studies[83]; three articles reported some consideration about the robustness of meta-analysis results after including or excluding non-IPD studies (e.g. whether the estimate effect of the PF is changed after including non-IPD studies[81]). Rovers et al.[87] stated that '*six out of ten studies were included in the meta-analysis, and the four excluded trials would have not changed the results of meta-anlayses'*. Only one study combined IPD and non-IPD studies to see the effect on the PF; for instance Thakkinstin et al.[81] stated that '*IPD was partially reconstructed using summary data in three non-IPD studies, and pooled with IPD studies using logistic regression*'.

2.4.7 Limitations and Challenges of an IMPF

Whilst reading the 20 IMPF articles, I recorded any pitfalls or challenges reported by the authors about the process of conducting IMPF projects (Figure 2.8); the three most common problems reported were: (i) unavailability of IPD for some studies (this may lead to the problem of publication and availability bias (see section 2.4.6 and chapter 7), (ii) different methods of measurement of the PF across studies and (iii) missing data.

Table 2. 7: Details about considering non-IPD studies and selection bias

Study ID	First Author	Was there a formal assessment of whether the included/selected IPD studies were a potential biased set of all available studies (akin to publication bias)?	If yes, what methods were used and what was concluded?	How many studies were asked for their IPD?	How many studies provided their IPD?	Was IPD obtained from all studies desired?	If no (or unclear): was the number of patients in the non-IPD studies given (or % missing)?	Were the number of events in the non-IPD studies given (or % missing)?	Were details given as to the robustness of meta-analysis results to the inclusion/exclusion of non-IPD studies? If yes, briefly state what was said?	Was a method used to combine IPD and non-IPD studies? If so, what was it?
1	Butcher	No	Not relevant	Not relevant	9	Not relevant	Not relevant	Not relevant	Not relevant	Not relevant
2	McHugh	No	Not relevant	Not relevant	11	Not relevant	Not relevant	Not relevant	Not relevant	Not relevant
3	Murray	No	Not relevant	Not relevant	11	Not relevant	Not relevant	Not relevant	Not relevant	Not relevant
4	Maas	No	Not relevant	Not relevant	11	Not relevant	Not relevant	Not relevant	Not relevant	Not relevant
5	Mushkudiani	No	Not relevant	Not relevant	11	Not relevant	Not relevant	Not relevant	Not relevant	Not relevant
6	Van Beek	No	Not relevant	Not relevant	7	Not relevant	Not relevant	Not relevant	Not relevant	Not relevant
7	Koopman	No	Not relevant	6	5	No	No	No	No	No
8	Thakkinstian	No	Not relevant	13	5	No	No	No	Yes, but despite there were still no statistically significant effects	Yes, IPD was partially reconstructed using summary data in 3 non-IPD studies, and pooled with IPD studies using logistic regression
9	MeRGE2	Yes	Inspecting funnel plot, and calculating the I^2 statistic and a test for heterogeneity	17	12	No	Yes	No	No	No
10	Yap	No	Not relevant	4	4	Yes	Not relevant	Not relevant	Not relevant	Not relevant

Continue on the next page

Continue from previous page

Study ID	First Author	Was there a formal assessment of whether the included/selected IPD studies were a potential biased set of all available studies (akin to publication bias)	If yes, what methods were used and what was concluded?	How many studies were asked for their IPD?	How many studies provided their IPD?	Was IPD obtained from all studies desired?	If no (or unclear): was the number of patients in the non-IPD studies given (or % missing)?	Were the number of events in the non-IPD studies given (or % missing)?	Were details given as to the robustness of meta-analysis results to the inclusion/exclusion of non-IPD studies? If yes, briefly state what was said?	Was a method used to combine IPD and non-IPD studies? If so, what was it
11	Rovers	No	Not relevant	19	6	No	No	No	Yes, 6 out of 10 studies included in the meta analysis, and the 4 excluded trials would have not changed the results of the meta-analysis.	No
12	Trivella	No	Not relevant	38	18	No	No	No	No	No
13	MeRGE1	No	Not relevant	32	18	No	Yes	No	No	No
14	Noordzij	No	Not relevant	15	9	No	No	No	No	No
15	Lanterna	Yes	Funnel plot for asymmetry and regression asymmetry test	8	7	No	Yes	Yes	No	No
16	Downing	No	Not relevant	6	4	No	Yes	No	Yes, they compare the meta-analysis result from the 4 IPD studies with the published results from the 6 studies	No
17	Sylaja	No	Not relevant	Unclear	4	Unclear	Not relevant	Not relevant	Not relevant	Not relevant
18	Schaich	No	Not relevant	8	8	Yes	No	No	No	No
19	Warkentin	No	Not relevant	8	7	No	No	No	No	No
20	Goetz	No	Not relevant	Unclear	11	Unclear	Non-IPD studies not mentioned	Non-IPD studies not mentioned	Non-IPD studies not mentioned	Non-IPD studies not mentioned

End of Table 2.7

Figure 2.8: Summary of the challenges facing researchers conducting an IMPF

Identifying all relevant studies

- Unavailability of IPD in some studies
- Time-consuming and costly nature of obtaining, cleaning and analysing the IPD.

Issues within individual studies

- Dealing with skewed continuous variables and possible outliers.
- Inability of IPD to overcome deficiencies of original studies, such as being retrospective rather than prospective, being too small for a multivariable analysis, missing important confounders, missing participant data or being of low methodological quality, etc.
- How to assess the quality of studies identified
- Dealing with the continuous variable if the linearity assumption is not achieved.

Heterogeneity between studies

- Different definitions of disease or outcome.
- Different participant inclusion and exclusion criteria.
- Different methods of measuring for the same prognostic factor.
- For survival data different lengths of follow-up
- Factors measured at different points in time or at different stages of disease across studies.
- Different (or out-dated) treatments strategies, especially when a mixture of older and newer studies is combined.
- Insufficient information about treatment for some of the studies.

Statistical issues for meta-analysis

- Missing data, including: missing factor values and outcome data for some participants within a study, and unavailable factors in some studies.
- Inability to adjust prognostic effects for a consistent set of adjustment factors in each study
- Imposed choice of cut-off levels when individual studies categorise their continuous variables and/or categorise their continuous outcomes in their provided IPD
- Difficulty in using a continuous scale for continuous factors in meta-analysis when some studies give IPD give values on a continuous scale and others do not (e.g. see Rovers et al.[97])
- Considering whether it is sensible and/or possible to investigate differential prognostic effects in subgroups
- Potential for study-level confounding when assessing whether study covariates (e.g. year of publication) modify the prognostic effect.
- Difficulty of interpreting summary meta-analysis results in the presence of heterogeneity across studies.
- When and how to account for clustering (‘one-step’, ‘two-step’ or just treat as all one study)
- How to combine IPD and non-IPD studies.

Assessment of potential biases

- Potential for publication bias and availability bias
- How to assess the robustness of IPD meta-analysis results to the inclusion/exclusion of studies only providing summary data; and how to combine IPD studies with summary data studies

Different Methods of Measurement

If the PF has different units of measurement across studies, there is difficulty in combining the IPD studies for the PF. Of the 20 IMPF articles, seven articles reported the problem of different methods of measurement[72 77 78 81 85 86 93]; three of these seven articles explained the causes of different methods of measurement of the PF across their IPD studies and explained how they limit and cope with this problem[72 78 86]. For example, Trivella et al.[72] state that 'there are two main methods for measurement' - so they do separate analyses for each one.

Even when the same method was used to count microvessel density (i.e., Chalkley vs. all vessels), individual laboratories used very different procedures for measurement of microvessel density. Four of the seven articles that reported the problem of different methods of measurement did not attempt to limit or cope with this problem[81 85 93]. For instance, Thakkinstian et al.[81] stated that **"***Since the IPD meta-analysis is a retrospective collaboration, it is difficult to get clinical variables that have been assessed and measured using similar methods across all studies"*. So it seems they are aware of it, but haven't been able to address it.

Missing Data

Missing data occurs when no data are available in PF or outcomes; in particular there are three types of missing data in IMPF articles; first, the missing variables occur when there is no data provided for a certain PF or confounding factors in some studies of the IPD studies provided. For example, in Look et al[14]. for the 15 IPD dataset, the PAi variable was not available in some of these IPD studies. Second, missing values occurs when some values are missing in a certain PF or confounding factors. For example, in Look et al[14]. across the 15 IPD dataset, there are 68 missing patients values for uPA across studies. Third, missing outcomes occure when there is some missing values of the outcome for some patients in the

study. All of the twenty articles reported at least one of these missing data problems. Nine of these twenty articles reported statistical methods to limit the problem of missing data; three of these nine articles directly mentioned to an 'Imputation Analysis', and gave a brief information about the methods of doing it[74 77 83]. Three of nine articles report how they overcome missing data. One of them reported 'missing value analysis function' whereby they imputed the missing data per trial using a linear regression method (e.g. Koopman et al.[82]) ; and in the remaining two articles the authors did not directly report to the statistical methods, they just reported that they imputed the data (e.g. missing outcomes, missing patient level) without any further information[75 78 79].

2.5 Discussion

Meta-analysis of IPD is well-known to be the "gold-standard" approach to combining evidence across multiple studies. This review has shown that the number of IPD meta-analysis articles have increased dramatically over time. Further, IMPF are also achievable and I identified 48 currently published. An in-depth evaluation of the 20 most recent IMPF revealed their methods and reporting are often sub-optimal, and there are numerous challenges for the IMPF approach, despite their numerous advantages over aggregated data.

2.5.1 Review Limitations

Though I used a systematic review that involved a wide search strategy to identify general IPD meta-analysis articles and IMPF articles, there is still possibility of missing some IPD articles and IMPF articles. Searching for and classifying IMPF articles was also not trivial; in particular, due to different types of prognosis studies, and the lack of taxonomy that exists in the field.

Another limitation is that the information extracted from the 20 recent IMPF articles is dependent on reporting standards therein, and so apparent deficiencies within an IMPF

project may be confounded by poor reporting. For example, only 3 IMPF articles referred to a protocol for their project, but this does not necessarily mean a protocol did not exist in the other 17 articles. Thus apparent gaps in study conduct may simply relate to gaps in study reporting. This is particularly important given I only elicited information directly available in the published IMPF article, and did not utilise other reports (e.g. protocols, statistical analysis plan) or contact authors for information directly. However, it is unlikely that the aforementioned issues would alter the main findings about the trend in IPD articles over time; the often sub-optimal methods and reporting in IMPF and the main challenges facing those conducting IMPF.

It is clear that IMPF articles still have many challenges. In this thesis, I tackle some of these challenges; at the end of the thesis (chapter 8), I provide recommendations for how IMPF projects can be improved.

2.5.2 Motivation for Further Research in this Thesis

There are many challenges facing statisticians, working within an IMPF, as identified during the review in this chapter (Figure 2.8), and these motivate the rest of my thesis. Missing data (e.g. missing variables, missing values and missing outcomes) is one of the most common problems in IMPF articles. From the 20 IMPF articles, some authors deal with this problem by omitting the missing values and excluded the variables that are missed in some IPD studies; other authors limit this problem by using the imputation analysis to estimate the missing data. But a clear strategy for dealing with missing data in IPD meta-analysis is not yet established. Another issue is that some of the IMPF articles did not account for clustering of patients within studies which might impact upon the accuracy of their results and this needs to be examined. Also is the choice between one-step and two-step meta-analysis approaches important?

Handling different methods of measurement of PFs across studies is another substancial problem not currently solved, and the strategy for modeling continuous variables and dealing with non-linearity in meta-analysis is not established. Thus, there is an array of statistical problems that needs to be tackled to ensure IMPF produces reliable, evidence-based results for PFs.

In subsquent chapters I tackle some of these issues further; in particular chapter 4 and 5 compares one-step versus two-step approaches, chapter 6 considers analysis of continuous PFs, and chapter 7 considers publication bias and availability bias. Before that I illustrate the issues using a real dataset in chapter 3.

CHAPTER 3

IPD VERSUS AGGREGATED DATA: AN EXAMPLE OF THE ADVANTAGES AND PROBLEMS

3.1 Introduction

This chapter aims to show through simple analyses and empirical evidence why a meta-analysis of aggregated data is limited severely for PF studies, and why an IPD meta-analysis is preferred. I will focus on the assessment of uPA as a PF in patients with breast cancer, see chapter 1, section 1.3, Table 1.2. First, I will highlight the problem of an aggregated data review as conducted by Look et al.[98] using IPD from 15 studies. I will demonstrate how IPD allows more reliable and clinically meaningful answers. Finally, I will demonstrate that IPD does not solve all the problems of a meta-analysis of PFs by highlighting some of the issues for IMPF articles identified in chapter 2.

3.2 Aggregate Data of Meta-analysis of uPA in Breast Cancer

The aggregate data approach to meta-analysis of PFs was introduced and exemplified in chapter 1, section 1.6, where the disadvantages of the approach were also noted. Some of the issues are apparent in an aggregate data meta-analysis of urokinase-types plasminogen activator system (uPA) in breast cancer by Look et al.[98]The uPA system comprises at least four proteins, urokinase-type plasminogen activator (uPA), its membrane-bound receptor (uPAR) and two plasminogen activator inhibitors (PAI-1 and PAI-2). These proteins have been suggested to be PFs in breast cancer. The balance between the expression levels of the various component of the uPA-system could determine whether active proteolysis and invasion takes place and whether cell migration will be affected by non-proteolytic or proteolytic mechanisms or by both. Look et al.[14] reviewed the literature for the uPA system

and relapse-free survival in breast cancer patients, and clinical important subgroup therein. The authors identified 41 relevant studies from which aggregated data were extracted regarding the study results; in particular, the univariate and multivariate hazard ratio results regarding the prognostic effect of the uPA system. For simplicity, in this section the focus is on a relation between uPA as a PF and relapse-free survival in breast cancer patients. The other factors such as uPAR, PAI-1 and PAI-2 are excluded.

Table 3.1 shows the aggregate data that could be extracted for uPA in those 7 studies reporting the relation between relapse-free survival and uPA for all patients. This table is taken from Look et al. [98] It can be seen from Table 3.1 that there are missing data for some studies such as the number of relapses during follow up, the univariate hazard ratio (i.e. the hazard ratio from a survival model just including uPA as a covariate) and the multivariable hazard ratio (i.e. the hazard ratio from a survival model including uPA and also other PFs). In addition, the variables included in the multivariable model differ from study to study, which is likely to cause heterogeneity in the adjusted uPA results across studies. Furthermore, there is a large difference in the choice of cut-off level among studies; for example, in one study the cut-off level for uPA is 0.62 and in another it is 4.45, which suggests no consistent strategy for determining the cut-off level and potentially different methods of measurements from study to study.

Another limitation of an aggregate data approach is to examine the ability to look at uPA as a PF in subgroups of patients. For example, one may wish to look at the prognostic effect of uPA in those group defined by a negative and positive hormone receptors status. Only 3 studies out of 41 considered this (Table 3.2), and only one of these provided the unadjusted

Table 3. 1: uPA and their relation with relapse-free survival in all patients[98]

Study reference number	Number of patients included	Cut-point	Fraction of tumors with levels above the cut-point	Number of Relapses during follow up	Follow up in months	Unadjusted P-value	Unadjusted hazard ratio (uHR)	Adjusted P-value	Adjusted hazard ratio(aHR(95% CI))	Other PFs[1]
uPA										
20	229	3.0	39	48	30	<0.0001	3.22	0.0002	3.0(1.7-5.5)	LN,PgR,PAI-1
17	184	0.81	23	?	83	0.0008	?	0.003	2.3(1.3-4.0)	LN,ER
42	688	0.62	33	173	42	0.0002	?	0.004	1.8(1.2-2.8)	LN,TAM,S-phase,PgR
46	429	4.45	50	201	61	0.0023	?	0.17	?	meno, LN, T, grade, PAI-1, HRS
26	314	0.52	32	91	84	0.06	?	?	?	Na
45	226	0.3	50	50	60	0.014	2.04	0.044	1.89(1.01-3.50)	age,LN,ER,adjCT
29	629	1.15	32	253	48	<0.0001	?	0.01	1.46(1.1-1.95)	age/meno,LN,T,PAI-1,ER,PgR,PS2,CD

[1] age/menopausal status: age and menopausal status, LN:Lymph-node status, T:tumor size, HRS:hormone-receptor status, TAM:adjuvant tamoxifen treatment, adjCT:adjuvant chemotherapy,CD:cathepsin D, ? Not shown, Na: not available.

Table 3.2: uPA and its relation with relapse-free survival in hormone receptor or ER -negative and positive patients[98].

Study reference number	Number of patients included	cut-point	Fraction of tumors with levels above the cut-point	Number of Relapses during follow up	Follow up in months	Unadjusted P-value	Unadjusted hazard ratio (uHR)	Adjusted P-value	aHR(95% CI)	Other PFs
Negative Hormone-receptor										
20	59	3.0	?	21	30	<0.21	1.79	?	?	na
13	70	10	63	?	64	ns	?	?	?	na
28	238	1.15	36	113	48	?	?	0.11	1.38(0.98-2.05)	age/meno,LN,T
Positive Hormone-receptor										
20	170	3.0	?	28	30	0.0002	3.72	?	?	na
13	96	10	41	>14	64	<0.001	?	?	?	na
28	406	1.15	29	138	48	?	?	<0.0001	2.76(1.96-3.87)	age/meno,LN,T

[2] age/menopausal status: age and menopausal status, LN:Lymph-node status, T:tumor size, HRS:hormone-receptor status, TAM:adjuvant tamoxifen treatment, adjCT:adjuvant chemotherapy,CD:cathepsin D, ? Not shown, Na: not applicable.

hazard ratio (uHR) and adjusted hazard ratio (aHR) for each group defined by hormone receptor.

All these issues make a meta-analysis of aggregate data hard to justify here. There is much missing data, and interpretation is severely limited by heterogeneity. Thus, meta-analysis is unlikely to be reliable or clinically meaningful. One could try to utlise the methods of Parmar et al[18], to obtain the estimates where currently missing, but Riley et al.[22] and Kyzas et al[7], show this is problematic for PF studies due to poor and selective reporting, and it would not solve most problems such as inconsistent cut-off and variable selection of adjustment factors.

3.3 IPD Meta-analysis of uPA in Breast Cancer

Look et al.[14] seek to overcome the aggregate data problems described in section 3.2 by obtaining IPD from 18 collaborating centers in the field and they have kindly provided me with the IPD from 15 of these centers. I will now use this to demonstrate the advantages of having the IPD over the aggregated data approach[56 61 94 99]. An IPD meta-analysis of PFs can be tackled using a 'one-step' or 'two-step' approach[66 94 100 101], but here I use the 'two-step' approach as this most closely reflects what the aggregate data approach is trying to achieve[99]. In the 'two-step' approach, the IPD are first analysed separately in each study using an appropriate statistical method for the type of data being analysed. For example, for time until the events occur, the Cox proportional hazard model might be fitted. This produces aggregated data for each study, such as the PF effect estimate and its standard error; then a fixed-effect or random-effects meta-analysis method is used to pool the data across studies to estimate the summary hazard ratio (see section 1.7).

3.3.1 Advantages of using the IPD Approach for uPA

(1). Increased number of studies and events

The number of the studies in my IPD breast cancer data was 15, comprising 7435 patients and 2645 events. Though the total number of the studies in aggregate data review was 41 in Look et al.[98] there were only 7 studies with 2699 patients and 816 events that provided aggregated data

PF results for uPA for the all breast cancer patients, and within these 7 studies there are missing estimates. This indicates that the number of studies, patients and events is larger in IPD compared to aggregated data approach. Also, if the data are classified according to confounding factor of hormone receptor, in the aggregated data approach for Look et al.[98] only 3 studies with 869 patients and 314 events (134 and 180 events for negative and positive hormone receptor respectively) were available. However, in the IPD approach, the whole 15 studies was available comprised 7435 patients with 2645 events (622 and 2023 events for negative and positive hormone receptor respectively).

(2). Perform meta-analysis with uPA on original scale

In the IPD, uPA was available on its original scale in each study[69 102], so I chose to maintain this scale and not dichotomize, to maximise statistical power. This is not possible using the aggregate data approach, as aggregated data results are presented with uPA analysed on dichotomized scale with inconsistent cut-off points across studies. I thus applied a two-step IPD meta-analysis with uPA on its original scale. In the first step I fitted a Cox model to each study separately ($i = 1\ to\ 15$), which can be illustrated mathematically as follows:

$$\lambda_i(t) = \lambda_{0i}(t) \cdot \exp[\beta_i \cdot uPA] \qquad (3.1)$$

where uPA is a continuous PF. In the second stage, I fitted a random-effects meta-analysis model to estimate the effect size of the PF, $\hat{\beta}_i$, as follows:

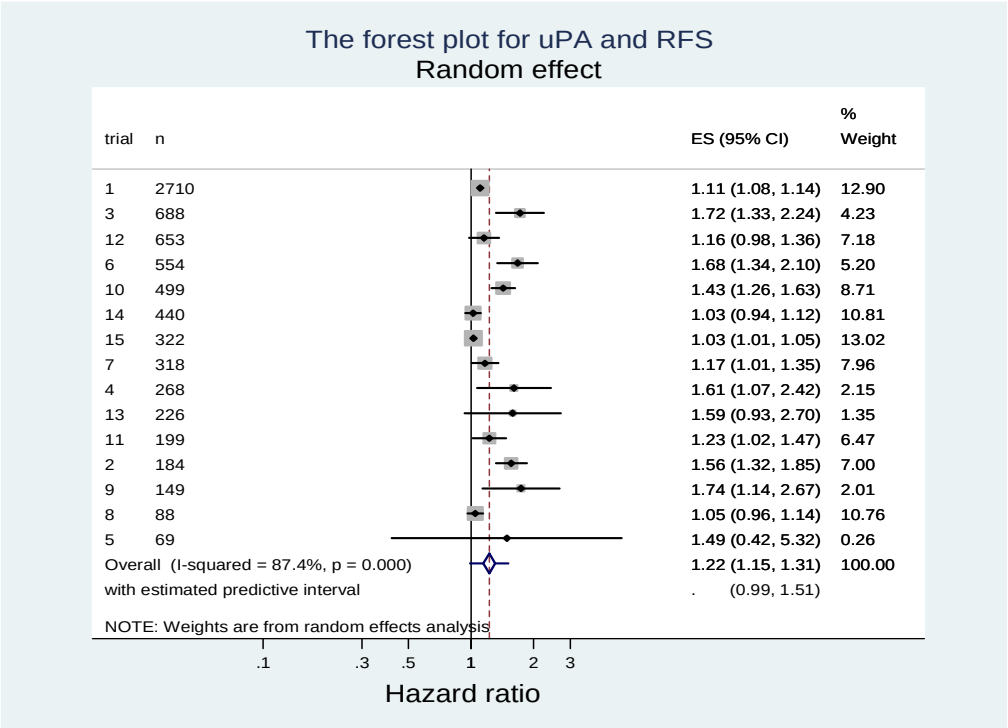
$$\hat{\beta}_i \sim N(\beta_i, var(\hat{\beta})) \qquad (3.2)$$

$$\beta_i \sim N(\beta, \tau^2)$$

where β is the average log hazard ratio across studies, and τ^2 is the between study heterogeneity in log hazard ratio. Figure 3.1 shows a forest plot of this IPD meta-analysis. It can be seen that there is large heterogeneity as $I^2 = 87.4\ \%$; this means 87.4% of the total

variability among effect sizes is caused by between study heterogeneity. The pooled hazard ratio is 1.22 which means that the mean effect (average of hazard rate) increases by 22% for every one unit increase in uPA; with 95% confidence interval for the average hazard ratio lies between 1.15 and 1.31. Thus it does not include one, and indicates a statistical significant relationship between uPA and RFS on average. Thus, even despite the heterogeneity, there is still strong evidence that uPA has prognostic value in an individual study setting. The estimated between-study variance of the true hazard ratio, τ^2 , is 0.0085.

Figure 3.1: A forest plot for uPA and RFS from the 15 IPD studies for the breast cancer patients.



(3). Examine prognostic factors in subgroup of patients

Another advantage of IPD is to look at the prognostic effect in specific subgroup of patients[103]. For example, one may wish to look at the effect of uPA in subgroup defined by hormone receptor status (e.g. negative or positive hormone receptor). This may lead to more specific recommendations about the role of uPA in individual patient, and help reduce between study

heterogeneity. As discussed in section 3.2 this was very difficult using aggregate data, due to the poor reporting of subgroup results, see Table 3.2.

In each study, I fitted Cox proportional hazard model for each hormone receptor group separately, the negative hormone receptor group model is as follows:

$$\lambda_{s_{0i}}(t) = \lambda_{s_{0i}}(t) \cdot \exp[\beta_{0i} \cdot uPA] \tag{3.3}$$

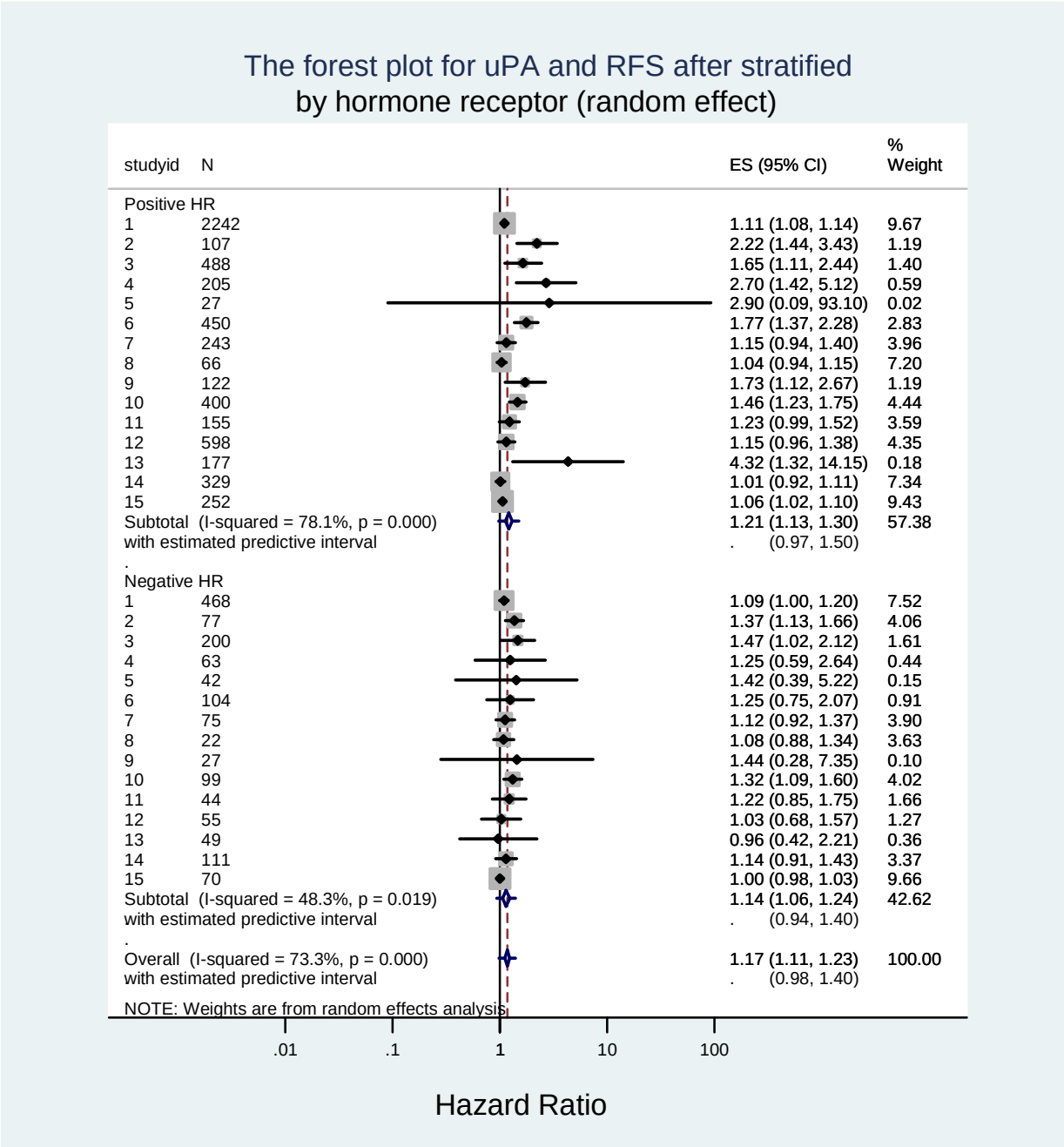
where s_0 refers to the negative hormone receptor group. The positive hormone receptor model in each study is:

$$\lambda_{s_{1i}}(t) = \lambda_{s_{1i}}(t) \cdot \exp[\beta_{1i} \cdot uPA] \tag{3.4}$$

where s_{1i} here refers to the positive hormone receptor, for each study group separately. I then performed a random-effects meta-analysis of the $\hat{\beta}_{0i}$ or $\hat{\beta}_{1i}$ estimates , as in equation 3.2.

Figure 3.2 shows the subgroup meta-analysis results obtained. In positive hormone receptor $I^2 = 78.1\,\%$, which means that 78.1% of the total variability among effects sizes is due to between study heterogeneity. The estimated variance of the true effect size of the positive hormone receptor group τ^2 is 0.0101. The estimated average hazard ratio for a 1-unit increase in uPA in the patients who had a positive hormone receptor is 1.21; this means that the hazard increases by 21% for every 1-unit increase in uPA for the patients with a positive hormone receptor. The 95% confidence interval is between 1.13 and 1.30; it does not include one indicating a statistically significant association between uPA and RFS for patients who have positive hormone receptor.

Figure 3.2: A forest plot for uPA and RFS from the 15 IPD studies for breast cancer patients after grouped by hormone receptor.



In negative hormone receptor, I^2 = 48.3% , which indicates that there is still heterogeneity across studies, but the proportion of heterogeneity for negative hormone receptor patients is much less than for positive hormone receptor. The estimated variance of the true effect size for negative hormone receptor group τ^2 is 0.0072. The estimated average hazard ratio is 1.14

slightly less than the average hazard ratio estimate for positive hormone receptor, and the 95% confidence interval, for negative hormone receptor is between 1.06 and 1.24.

The 95% confidence interval for the average hazard ratio is again entirely above 1. So, these analysis have shown that uPA is, at least on average, prognostic for RFS in both subgroup defined by hormone receptor status. A test of a subgroup effect (i.e. a difference in hazard ratio between hormone receptor groups), using a meta-analysis model was not significant. This can be seen by the largely overlapping confidence intervals for the average of the meta-analysis results.

(4). Adjust analyses for a consistent set of other PFs

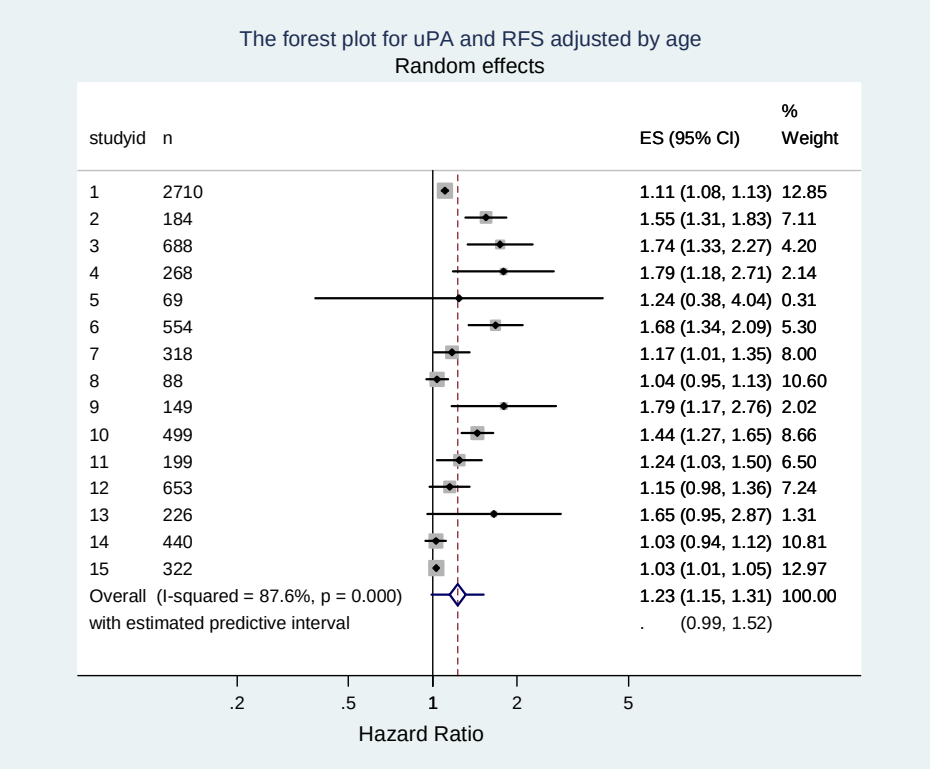
Another advantage of IPD is to look at PFs whilst adjusting consistently for other variables. To demonstrate this, I will look at the effect of uPA in all patients after adjusted by age. First, in each study separately I fitted a Cox proportional hazard model for uPA in all patients after adjusting by age. The model is written below:

$$\lambda_i(t) = \lambda_{0i}(t) \cdot \exp[\beta_{1i} \cdot uPA + \beta_{2i} \cdot age] \quad (3.5)$$

Where β_{1i} and β_{2i} refer to the ln(HR) for a 1-unit increase in uPA and age respectively. Then I fitted a random-effects meta-analysis of the $\hat{\beta}_{1i}$ estimates obtained.

Figure 3.3 shows the forest plot for this meta-analysis. The estimated between study variance of the τ^2 is 0.0088. The results again show that the uPA is an independent PF even after adjusting for age. Such a finding was not available without IPD.

Figure 3.3: A forest plot for uPA and RFS from the 15 IPD studies for breast cancer patients after adjusted by age



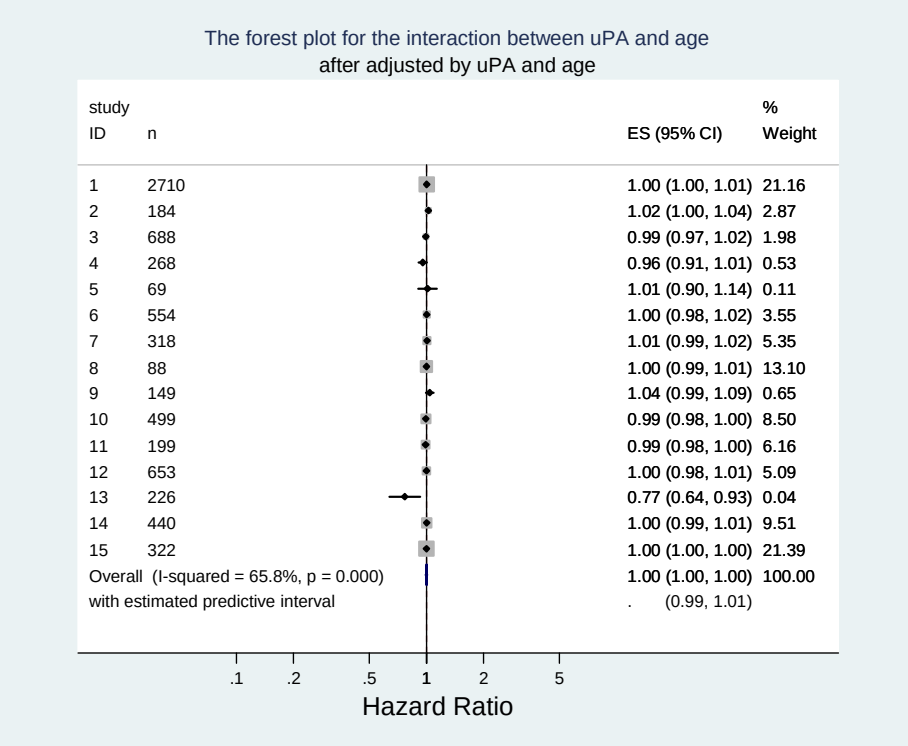
(5). Assess the interaction between two PFs as an additional PF

Another advantage of the IPD approach is to measure the effect of the interaction between two PFs. For example, I now assess whether the interaction between uPA and age is an independent PF itself, over uPA and age themselves. I fitted the cox proportional hazard model as follows in each study:

$$\lambda_{1i}(t) = \lambda_{0i}(t).\exp\left[\beta_{1i}.uPA + \beta_{2i}.age + \beta_{3i}.(\text{uPA} * \text{age})\right] \quad (3.6)$$

Then I fitted a random-effects meta-analysis to the $\hat{\beta}_{3i}$ estimates.

Figure 3.4: A forest plot for the interaction between uPA and age after adjusted by uPA and age from 15 IPD studies for breast cancer patients



It can be seen from Figure 3.4 that the hazard ratio is one which indicates that there is no significant effect for the interaction between uPA and age on risk of mortality (P-value > 0.05); and there is heterogenity across studies (65.8 %). Note that without the IPD, you are reliant on the studies themselves reporting the interaction estimates and their standard errors, but this rarely happens in practice and was not available in aggregated data approach here.

The five illustrated advantages show some of the key advantages of the IPD approach over the aggregate data approach, using more patients and events to consider PFs in subgroups and having adjusted for other factors. Other advantage exists, such as modeling non-linear trends (see chapter 6). But, as identified in chapter 2, there are still numerous challenges for conducting the IPD approach. I now illustrate some of these:

(1). IPD not available for all studies

Though the number of the IPD studies for Look et al.[14] was 18 studies, only 15 IPD studies were provided to me. The other three studies were not provided to me as the original authors refused, as they also wanted to use their data for methodological research. This reflects sensitive issues around data sharing, and was seen in my review in Chapter 2, where many IMPF projects could not obtain all the IPD required.

(2). Missing data

There are missing data in the Look et al.[14] IPD dataset. There are missing variables in some studies. For example, some PFs are entirely missed in some IPD studies (i.e. PAI-1), and some PFs have missing values for some patients in some studies (e.g. in total there are 68 missing patients values for uPA across studies).

(3). Different methods of measurements

Look et al.[14] report that the uPA was measured using either an Enzyme-linked immunosorbent assay (ELISA) kit, a Luminometric Immuno Assay (LIA) kit, or a protein assay kit. Of the ELISA kits, there were 3 different varieties used across studies; of the LIA kits, there were at least 2 different varieties; and of the protein assay kits, there were 3 varieties. Look et al.[14] note that this causes heterogeneity in uPA values across studies, causing the median and range of uPA values to vary from study to study. This will undoubtedly be contributing to the unexplained heterogeneity in my illustrated meta-analyses. Look et al.[14] try to overcome the problem by transforming uPA values in each study onto a rank score, but I do not consider this here as their approach has not been validated in a methodological paper yet, and the transformation does not aid clinical interpretability.

(4). Model assumptions

Another advantage of having IPD is the ability to check model assumptions. But the issue arises about what to do if model assumptions fail in some studies? I checked the assumption of proportional hazards for each of the 15 IPD studies separately, and found that in 14 out of 15 studies the assumption holds for RFS and OS, with a P-value for uPA from schoenfeld residuals test > 0.10. This generally confirms that Cox proportional hazard assumption is a suitable model. One could do a sensitivity analysis including and excluding the 'fail' study, and this does not change conclusions here. I also checked the Cox proportional hazrd assumption for uPA transformed in a rank score, whick look et al[14]. used to overcome different methods of measurement, I found that the proportional hazard assumption does not hold in 3 and 4 studies for RFS and OS respectively, which indicates that the Cox proportional hazard model is not suitable in this case, and so further reasons are to why the rank trandformation appears inappropriate.

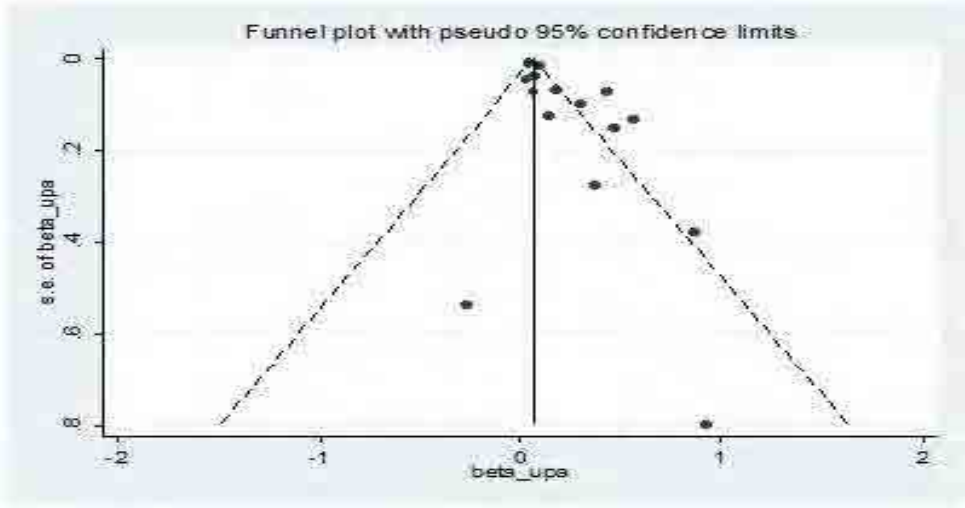
(5). Availability bias and publication bias

Another concern for the IPD approach is that the set of available IPD studies may be influenced by publication bias or availability bias, where the studies providing IPD are a biased set of all studies[14 22]. One way to assess such bias is by funnel plot which is a graphical tool for displaying the relationship between study effect size and its precision[104]. If there is no bias, the assumption is that the graph display a funnel shape with more variability around the average value across studies with low precision than across studies with high precision; asymmetry in the plot may indicate availability or publication bias, though it is known that given heterogeneity asymmetry may also occur naturally[105].

Figure 3.4 shows the funnel plot for the ln(HR) estimates for the univariate uPA in the 15 IPD studies. It can be seen from the graph that there does appear to be asymmetry with more studies

to the right of centre, than the left. Availability and publication bias may be causing this Egger's test of asymmetry to be highly significant (P=0.002), see Egger et al[106]. Thus, if availability or publication bias is causing the asymmetry, then the PF result is likely to be biased in favour of uPA being prognostic here. This issues is investigated again in chapter 7. Of course, the asymmetry may also be due to other reasons, such as the heterogenity covered by different methods of measurement

Figure 3.4: Funnel plot for assessment of availability bias in the 15 IPD studies in breast cancer, in relation to the meta-analysis of uPA for RFS



3.4 Discussion

In this chapter I have produced some very simple examples to demonstrate the problem of the aggregate data approach to meta-analysis of PFs and the advantages and the challenges of the IPD approach, in the context of whether uPA is a PF in breast cancer. The works builds on the review in chapter 2, that showed IMPF articles are increasingly common but that there are many challenges therein.

In Jones et al.[99] the authors compared between IPD and aggregated data meta-analysis approaches for repeated measurments models. The authors reported for two problems for meta-analysis of aggregated data: (i) the lack of the information about the correlation coefficients; (ii)

the effect of the missing data at the patent-level. This generally confirms my results here that there is a lack for the information of the aggregated data across studies. In Lyman et al.[57]the authors investigate the strengths and limitations on aggregated data. The authors concluded that IPD approach has many advantages over aggregated data. However IPD approach still have some limitations such as the cost and the time required to collect these data. This is again compatible with my conclusion. Also, in Stewart et al.[107] the authors investigated whether there is a difference between aggregated data and IPD data results, by applying on randomised trials of cisplatin-based therapy in ovarian cancer; the authors stated that '*The results of a meta-analysis of the literature alone may be misleading. Whenever possible, a meta-analysis of updated individual patient data should be done because this provides the least biased and most reliable means of addressing questions that have not been satisfactorily resolved by individual clinical trials*'.

In Berlin et al.[108] the authors compared between IPD and aggregated data approach by applying on 5 randomized trials; in particular, they investigate the benefits of anti-lymphocyte antibody induction therapy among renal transplant patients; they also considered the subgroup analysis to see whether the treatment might be more effective with a certain group of patients. The authors stated that '*Although patient-level analyses are not always feasible, for reasons of time and cost, we suggest that such analyses be used whenever subgroup analyses are deemed important.'* Further in Smith et al.[59] the authors concluded that when IPD are available, it improve the potential to investigate and explain heterogeneity across studies, and also the clustering across studies can be undertaken by using hierarchical model.. These two articles are consistent with my findings that the IPD data is better than aggregated data if subgroup analyses are required.

In Williamson et al.[109] the authors stated that '*Methods of estimating log(HR) and its variance and constructing graphical displays using aggregate data extracted from published survival curves can only be approximate and will often be constrained by the quality of such Figures. The limitations of using survival curves to estimate the log (HR) include survival curve inter-reader*

variability and the need to make some assumption about censoring'. And the authors also stated that *'A significant discrepancy would indicate that the aggregate data meta-analysis is unreliable and individual patient data are needed'*. This again indicates that the results from aggregated data are often unreliable, and IPD meta-analysis approach is in favour over aggregated meta-analysis approach.

In conclusion, IPD approach have many advantages over the aggregated data approach, however it still have some challenges and pitfalls (e.g. missing data, publication bias, nonlinear trend). Solving some of these problems forms the motivation for the remainder of this Ph.D thesis. In particular, in the next two chapters I develop one-step and two-step methods of meta-analysis and undertake a simulation study to investigate and generalize which model is the best. In addition, in chapter 6 I will examine the effect of nonlinear trends by using fractional polynomial models within one-step and two-step approaches. Then, in chapter 7 I will examine the extent of publication bias in IMPF articles.

CHAPTER 4

STATISTICAL MODELS FOR FIXED-EFFECT IPD META-ANALYSIS OF A SINGLE PROGNOSTIC FACTOR WITH A BINARY OUTCOME

4.1 Introduction

Chapter 2 identified many issues for an IMPF; one such issue is how best to perform meta-analysis, and which statistical method is best to use. In particular should the meta-analysis account for the clustering of patients within studies, and is a one-step or two-step method more appropriate? Recall that through my reviewing of the 20 IMPF articles in chapter 2, 14 of the 20 articles mentioned they accounted for the clustering of patients within studies in their analysis. Five used a 'one-step' method[80 81 85 88 93] with study as a covariate, seven used a 'two-step' method[72 74 75 77 79 83 90], one used both 'one-step' and 'two-step' approaches[89]; and one accounted for clustering but it was not clear whether a 'one-step' or 'two-step' was used[78]. In addition, five other articles used a 'one-step' method but did not account for clustering of patients within studies, essentially treating the IPD as if it is from one study. Indeed, it is clear that the statistical differences between one-step and two-step method is still unclear, as there is not a common used approach, and even accounting for clustering is not always considered important.

In this chapter, I develop a range of possible IPD meta-analysis models for examining a single PF for a binary outcome. These include one-step and two-step approaches, and those that do or do not account for clustering. Other studies have shown that non-linear models that do not account for important covariates produce estimates that are attenuated towards the null[110 111]. The 4 IPD studies of the TBI dataset that investigate whether there is an association between age as a PF and probability of death by 6 months (see Table 1-2) is used to illustrate the methods. Note that in this chapter I do not consider methods that account for heterogeneity in PF effect across

studies as there is no heterogeneity for the TBI dataset (see section 1.7.3). However, extension of the model to include random-effects is briefly considered.

4.2 Overview of Possible Fixed-effect IPD Meta-analysis Models for a single PF and Binary Outcome

Assume that there are $i=1$ to n independent IPD studies that each assesses the binary outcome of interest for n_i patients. Let y_{ik} be the binary outcome (1=dead, 0=alive) of patient k in study i and let x_{ik} be a PF, which could be continuous or binary[112]. Each IPD study thus provides y_{ik} and x_{ik} for the n_i patients in the study.

A range of different models are potentially available to meta-analysis such IPD (see Table 4.1). The difference in the model structure arises from using a one-step or two-step approach; accounting for unexplained residual variation due to the included factor; whether clustering of patients within studies are accounting for, and whether the correlation between $\hat{\propto}_i$ and $\hat{\beta}_i$ was utilised in meta-analysis. I will explain these models in details in the following sections. The models adapt those proposed by Riley et al.[112] for modelling baseline risk; here the focus is not on the baseline risk itself but on the PF effect.

4.3 Possible Two- step Models

In the 'two-step' approach, the IPD are first analysed separately in each study using an appropriate statistical method for the type of data being analysed[45 113-117]; for example, for a binary outcome a logistic regression model might be fitted. This produces aggregate data for each study, such as the PF effect estimate and its standard error; these are then synthesized in the second step using a suitable model for meta-analysis of aggregate data[55 60], such as one weighting by the inverse of the variance whilst assuming fixed or random PF effects across studies, see chapter 1. The two-step approach naturally accounts for clustering of patients within

studies by analysing each study seperately. For our situation the possible two-step models are given below, assuming a fixed-effect across studies.

4.3.1 Model 1: Fully Two-step Model (not applicable)

Model (1) is written as follows:

$$
\begin{aligned}
First\ step: \quad & logit(P_{ik}) = \propto_i + \beta_i x_{ik} + \varepsilon_{ik} \\
& \varepsilon_{ik} \sim N(0, \sigma_i^2) \\
Second\ step: \quad & \hat{\beta}_i = \beta + e_i \\
& e_i \sim N\left(0, Var\ (\hat{\beta}_i)\right) \\
& \hat{\beta} = \frac{\sum_{i=1}^{k} w_i \hat{\beta}_i}{\sum_{i=1}^{k} w_i}
\end{aligned}
\qquad (4.1)
$$

$\propto_i$ is the log-odds of event in study i for patients with$x_{ik} = 0$; β_i represents the change in log-odds of the event for one unit increase in x_{ik} . Also, x_{ik} is the PF for the patient k in study$i$. In the second step the pooled PF effect size, β across IPD studies is estimated by assuming a fixed-effect method as discussed previously. Note that w_i is the relative weight for each study i [See section 1.6].

In the first step the PF effect size, β_i, (i.e. log-odds ratio) is estimated by using a logistic regression model in *each* study separately; $logit\ (P_{ik})$ specifies a separate log-odds for each patient in each IPD study, Note that ε_{ik} has variance σ_i^2 , where σ_i^2 represents the variance in $logit(P_{ik})$; ε_{ik} is the residual error around $logit\ (P_{ik})$ for patient k in study i; this is akin to a linear regression, where you have residuals for each patient around the fitted line. However ε_{ik} is impossible to estimate here because ε_{ik} measures the within-study variance for each patient k in each study i, and the event for each patient just occurs once, so it is impossible to estimate the residual variance for each patient (as you do not know their $logit\ (P_{ik})$ value) and so σ_i^2 is not

identifiable; model 1 is thus not estimable. Note that model 1 is estimable if the outcome is continuous, and I did not consider that in my thesis, as the focus here is on binary outcome.

4.3.2. Model 2: Standard Two-step Model

Model 2 below is similar to model 1, except that in this model I assume that there is no residual variation, and thus remove the ε_{ik} as follows:

$$\begin{aligned} &First\ step: && logit(P_{ik}) = \propto_i + \beta_i x_{ik} \\ &Second\ step: && \hat{\beta}_i = \beta + e_i \\ & && e_i \sim N\left(0, Var\left(\hat{\beta}_i\right)\right) \\ & && \widehat{\beta} = \frac{\sum_{i=1}^{k} w_i \hat{\beta}_i}{\sum_{i=1}^{k} w_i} \end{aligned} \quad (4.2)$$

The parameters in model 2 are as defined as in model 1. This is now estimable.

4.3.3 Model 3: Advanced Two-step Model

However, it is possible for an intermediate model between model 1 and 2[112]. In model 3, I specify a distinct error term, $\varepsilon_{i_{(x_{ik})}}$, for each set of patients with the same value of x_{ik} in *each* IPD study. This allows for unexplained variation due to the covariate x_{ik} to be accounted for. Residual variation due to other unmeasured covariates is not considered; for the reasons outlined in section 4.3.1. Model 3 is written as follows:

$$\begin{aligned} &First\ step: && logit(P_{ik}) = \propto_i + \beta_i x_{ik} + \varepsilon_{i_{(x_{ik})}} \\ & && \varepsilon_{i_{(x_{ik})}} \sim N(0, \sigma^2_{i_{(x_{ik})}}) \\ &Second\ step: && \hat{\beta}_i = \beta + e_i \\ & && e_i \sim N\left(0, Var\left(\hat{\beta}_i\right)\right), \\ & && \widehat{\beta} = \frac{\sum_{i=1}^{k} w_i \hat{\beta}_i}{\sum_{i=1}^{k} w_i} \end{aligned} \quad (4.3)$$

The parameters here have the same interpretation as in model 1. $\sigma^2_{i_{(x_{ik})}}$ relates specifically to the remaining unexplained within-study variation in study i caused by the PF; in other words, it is the variance due to x_{ik} in study i that is not explained by the specification of $\beta_i x_{ik}$ in the model[112]. For example, if x_{ik} should be included as a quadratic term in truth, but only $\beta_i x_{ik}$ is included then there will be residual variation. However, the approach only works when x_{ik} is a continuous covariate, with multiple patients with exactly the same x_{ik} value as in the TBI example (it may also be possible for the categorical variable when the number of categorize need to be large); otherwise the residual variation is impossible to estimate as discussed in section 4.3.1. Note that $\varepsilon_{i_{(x_{ik})}}$ denotes the within-study error in study i for patients with the same PF value as patient k.

4.3.4 Model 4: Bivariate Fixed-effect Two-step Model

Models 1 to 3 did not account for the correlation between $\hat{\alpha}$ *and* $\hat{\beta}$ in the second step model, so the question is raised; *is it necessary to account for the correlation between* $\hat{\alpha}$ *and* $\hat{\beta}$ *in the two-step meta-analysis?* In other words *does the correlation between* $\hat{\alpha}$ *and* $\hat{\beta}$ *effect on the estimated pooled value of* $\hat{\beta}$ *in the two-step meta-analysis?* Bivariate fixed-effect meta-analysis model (BFMA) is used to consider within-study correlation[118-122], but before I explain the BFMA model I applied logistic regression model to one primary study of TBI dataset to see whether there is correlation between $\hat{\alpha}$ *and* $\hat{\beta}$ within a single study. I found that there was a large negative correlation (-0.938) between $\hat{\alpha}$ and $\hat{\beta}$ (when the estimated value of the PF increase, the intercept decrease). Thus this correlation can be accounted for by performing a BFMA model.

BFMA is a statistical model that synthesis $\widehat{\propto}_i$ and $\hat{\beta}_i$ together accounting for their correlation; model 4 can be written mathematically as follows:

$$First\ step: \qquad logit(P_{ik}) = \propto_i + \beta_i x_{ik}$$

$$Second\ step: \qquad \begin{pmatrix} \widehat{\propto}_i \\ \hat{\beta}_i \end{pmatrix} \sim N\begin{pmatrix} \alpha_i \\ \beta \end{pmatrix}, \delta_i\Big), \tag{4.4}$$

$$\delta_i = \begin{pmatrix} var(\widehat{\propto}_i) & cov(\widehat{\propto}_i, \hat{\beta}_i) \\ cov(\widehat{\propto}_i, \hat{\beta}_i) & var(\hat{\beta}_i) \end{pmatrix},$$

$$\propto_i \sim N(\propto, \tau_\alpha^2)$$

Where δ_i is the within-study covariance matrix. $\hat{\alpha}$ and $\hat{\beta}$ are estimated by using restricted maximum likelihood method[121 123]. Now $cov(\widehat{\propto}_i, \hat{\beta}_i) = -var(\widehat{\alpha_i})$ if x_{ik} is binary (see Appendix B1 for the proof of this). In this situation if $\tau_\alpha^2 = 0$ then $\hat{\beta}$ and $var(\hat{\beta})$ from the bivariate meta-analysis of model 4 are equivalent to$\hat{\beta}$, and $var(\widehat{\beta})$ from a standard univariate meta-analysis (as in step 2 of model 2), see appendix B2 for proof. This is perhaps why traditional meta-analysis methods do not worry about within-study correlation as where effect size relates to a binary covariate (e.g. a treatment), it makes no difference if the baseline risk ($\propto_i$) are simillar. However if $\tau_\alpha^2 \neq 0$ then $\hat{\beta}$ can differ if correlation is ignored (model 2) or accounted (model 4). This will be shown in section 4.5. Note that compared to previous models, model 4 makes an additional assumption that the $\propto_i$ are normally distributed about a mean$\propto$. Note also that if x_{ik} is continuous, $\hat{\beta}$ and $var\ (\widehat{\beta})$ can also differ between a BFMA utilising correlation and a univariate model.

4.3.5 Model 5:Bivariate Fixed-effect Advanced Two-step Model

Model 5 below is similar to model 4, except that in this model, we assume that there is also remaining residual variation due tox_{ik}. Model 5 is written as follows:

$$First\ step: \quad logit(P_{ik}) = \propto_i + \beta_i x_{ik} + \varepsilon_{i_{(x_{ik})}}$$

$$e_{i_{(x_{ik})}} \sim N(0, \sigma^2_{i_{(x_{ik})}})$$

$$Seond\ step: \quad \begin{pmatrix} \hat{\propto}_i \\ \hat{\beta}_i \end{pmatrix} \sim N\begin{pmatrix} \alpha_i \\ \beta \end{pmatrix}, \delta_i\Big),$$

$$\delta_i = \begin{pmatrix} var(\hat{\propto}_i) & cov(\hat{\propto}_i, \hat{\beta}_i) \\ cov(\hat{\propto}_i, \hat{\beta}_i) & var(\hat{\beta}_\iota) \end{pmatrix}, \qquad (4.5)$$

$$\propto_i \sim N(\propto, \tau^2_\alpha)$$

The parameters in model 5 are as defined as in model 3 and 4.

4.4 One-step Models

In the 'one-step' approach, the IPD from all studies are modeled simultaneously whilst potentially accounting for the clustering of patients within studies; this again requires a model specific to the type of data being synthesized, alongside appropriate specification of the meta-analysis assumptions[60 66 69 94 100 124 125] (e.g. fixed or random-effects across studies), see chapter (1). I now consider possible one-step models for meta-analysis of the prognostic effect of a covariate x_{ik} from IPD studies under a fixed- effects assumption.

4.4.1 Model 6: Ignoring the Clustering of Patients within-studies

Model (6) fits a logistic regression, by pooling all of the data together without considering the clustering of patients within-studies (considering all patients as if they came from one study).

$$logit(P_{ik}) = \propto + \beta x_{ik} \qquad (4.6)$$

This model specifies a single value for the intercept term (log-odds of the event for a patient with$x_{ik} = 0$) for all of the patients in all studies, and thus ignores the clustering of patients within-studies entirely (essentially assuming a constant baseline risk in each study); β indicates the change in log-odds of the event for one-unit increase inx_{ik}, and again this is assumed the

same (fixed) in each study. Note this model does not have a two step counterpart, as the two-step method naturally assumes a different $\propto_i$ in each study.

4.4.2 Model 7: Ignoring the Clustering of Patients within-studies and Accounting for any Remaining within-study Error across Patients with the Same PF Values

Model 7 is similar to model 6 in that it ignores clustering of patients within-studies, but it also accounts for any remaining within-study error across patients with the same x_{ik} values, that is not explained by the inclusion of βx_{ik} in the model; in this sense it is similar to the first step of model 3

$$logit(P_{ik}) = \propto + \beta x_{ik} + \varepsilon_{x_{ik}} \quad (4.7)$$

$$\varepsilon_{x_{ik}} \sim N(0, \sigma^2_{x_{ik}})$$

The parameters here have the same interpretation as in model (6), except that $\varepsilon_{x_{ik}}$denotes the within-study error in all IPD studies for patients with the same x_{ik} values; normally distributed about a common mean of zero and variance $\sigma^2_{x_{ik}}$. As for model 3, $\sigma^2_{x_{ik}}$ relates specifically to the remaining within-study variation due to x_{ik} that is not explained by the specification of βx_{ik} in the model.

4.4.3 Model 8: Accounting for the Clustering of Patients Within-studies by using Indicator Variables

While models 6 and 7 ignore the clustering of patients within studies, model (8) accounts for the clustering of patients within studies by including study indicators, that allow a separate intercept term for each study[126]. This model estimates the effect size of the PF across IPD studies as follows:

$$logit(P_{ik}) = \propto_i + \beta x_{ik} \quad (4.8)$$

Where α_i is the log-odds of event in study i when $x_{ik} = 0$ and β is the fixed PF effect size; it indicates the change in log-odds for a one-unit increase in x_{ik}.

4.4.4 Model 9: Accounting for the Clustering of Patients within-studies by using Indicator Variables and Accounting for any Remaining within-study Error across Patients with the Same PF Values

Model 9 is an amalgam of models 7 and 8. Model 9 is similar to model 8 in that it accounts for the clustering of patients within studies by using study indicator variables, and similar to model 7 in that it accounts for any remaining within-study error across patients with the same x_{ik} values for *each* IPD study

$$logit(P_{ik}) = \propto_i + \beta x_{ik} + \varepsilon_{i_{(x_{ik})}} \qquad (4.9)$$

$$\varepsilon_{i_{(x_{ik})}} \sim N(0, \sigma^2_{i_{(x_{ik})}})$$

The parameters here have the same interpretation as in model (8), with additionally $\varepsilon_{i_{(x_{ik})}}$ giving the within-study error in each study i for patients with the same x_{ik} value.

4.4.5 Model 10: Accounting for the Clustering of Patients within-studies by Placing a Random-effects on the Intercept term (Different Intercepts in each Study)

Model 10 is similar to model 8 in that it accounts for the clustering of patients within studies, but it achieves this differently. Whereas model 7 estimates a separate intercept for each study, model 10 places a random-effects on the intercept term to allow it to vary across studies according to a normal distribution with mean μ_α and variance σ^2_α

$$logit(P_{ik}) = \propto_i + \beta x_{ik} \qquad (4.10)$$

$$\propto_i \sim N(\propto, \tau^2_\alpha)$$

The parameters here have the same interpretation as in model 8. However, by assuming normality of the intercept value across studies, it makes a stronger assumption than Model 8 which does not impose a normal distribution relationship for the $\propto_i$.

4.4.6. Model 11: Accounting for the Clustering of Patients within Studies by Placing a Random-effect on the Intercept Term and Accounting for any Remaining Within-study Error across Patients with the Same PF Values

The final model, Model (11), is an extension of Model (10) to allow for residual variation across patients with the same value of x_{ik}, similar to model 3, 5 and 9.

$$logit(P_{ik}) = \propto_i + \beta x_{ik} + \varepsilon_{i_{(x_{ik})}} \tag{4.11}$$

$$\varepsilon_{i_{(x_{ik})}} \sim N(0, \sigma^2_{i_{(x_{ik})}})$$

$$\propto_i \sim N(\propto, \tau^2_\alpha)$$

The parameters here have the same interpretation as in model 9, and $\sigma^2_{i_{(x_{ik})}}$ are the remaining residual variation caused by x_{ik} that is not accounted for by the inclusion of βx_{ik} in the model. For simplicity here, the covariance between $\varepsilon_{i_{(x_{ik})}}$ and $\propto_i$ is assumed zero, but this can be relaxed if necessary.

4.5 Summary of all of the possible IPD models by using two-step and one-step approach

All of the possible models that explained in section 4.3 and 4.3 are summarized in Table 4.1. Again, the difference in the model structure arises from using a one-step or two-step approach; accounting for unexplained residual variation due to the included variable, whether the clustering of patients across studies are accounting for, and whether between $\hat{\propto}_i$ and $\hat{\beta}_i$ was utilized in meta-analysis.

Table 4.1: Summary of all of the possible IPD models by using two-step and one-step approach

Model No.	One-step or Two-step method	Account for clustering of patients within studies (if yes, how)	Account for remaining residual variation due to the included PF?	Accounting for correlation between $\widehat{\propto}_i$ and $\widehat{\beta}_i$ in the meta-analysis?
1	**Two-step**	Yes -by using two-step method	Yes 'fully' and not applicable	No
2	**Two-step**	Yes -by using two-step method	No	No
3	**Two-step**	Yes -by using two-step method	Yes 'partial'	No
4	**Two-step**	Yes -by using two-step method	No	Yes
5	**Two-step**	Yes -by using two-step method	Yes 'partial'	Yes
6	**One-step**	No	No	Yes
7	**One-step**	No	Yes 'partial'	Yes
8	**One-step**	Yes-by using indicator variables of the studies	No	Yes
9	**One-step**	Yes-by using indicator variables of the studies	Yes 'partial'	Yes
10	**One-step**	Yes-by placing a random intercept term (different intercept in each study)	No	Yes
11	**One-step**	Yes-by placing a random intercept term (different intercept in each study)	Yes 'partial'	Yes

4.5 Methods of Estimation

Note here that all one-step models naturally account for correlation between $\propto_i$ and β terms as they are estimated together. Model 2 and 4 (in two-step approach) and model 6, 8 and 10 (in one step approach) can be fitted classically by using maximum likelihood estimation. In contrast, model 3, 5 (two-step) and model 7, 9, 11 (one-step) are mixed models contain both fixed and random effects; the fixed effects are analogues to standard logistic regression coefficients and they can estimated directly by using maximum likelihood estimation. The random effects are not directly estimated but are summarized according to their estimated variances and covariances. The distribution of the random effects is assumed to be Gaussian, and an adaptive quadrative method is used to estimate the random effects for the model; this was undertaken by using the xtmelogit procedure in STATA[127 128].

4.6 Application to the TBI Data

4.6.1 Application for Continuous PF

I now apply models 2 to 11 to the TBI dataset to examine if age, x_{ik}, is a continuous PF for risk of six month mortality. Table 4.2 gives the results and I now discuss this.

Heterogeneity

Recall as in Chapter 1.7.3 there is no heterogeneity across studies as $I^2 = 0\%$ in this dataset, thus the fixed-effect approach is appropriate here. Though the aforementioned models can be extended to allow for heterogeneity for in the PF effect, this can be considered as future work (see chapter 8).

Two-step Models

Odds Ratio

As can be seen in Table 4.2, there is barely any difference in the pooled odds ratio estimate between models 2 to 5; as the pooled odds ratio is from 1.38 for the 4 models. All of the two-step models produce p-values <0.01, which indicate that there is a significant association between age and mortality rate (i.e. age is a PF).

Standard Error and Confidence Interval for $\hat{\beta}$

There is barely any difference for the confidence interval and standard error of $\hat{\beta}$ between models 2 to 5; however, the standard errors for models 3 and 5 are the highest standard errors[122], as these models account for the residual variance for the same value(s) of the PF in each study.

Table 4. 2: Meta- analysis results from possible IPD meta-analysis models to the TBI data when age is continuous.

Model type	Model number	Accounting for clustering	Accounting for residual variation due to x_{ij}	Accounting for correlation	Odds ratio	95% confidence interval for Odds ratio	Ln (odds ratio) and s.e.	95% confidence interval for Ln Odds ratio	Within study variance σ_i^2	P-value for pooled effect-size, $\hat{\beta}$
Two-step models	Model 2	Yes	No	No	1.38	1.29- 1.45	0.32 (0.030)	0.26 - 0.37	----------	<0.01
	Model 3	Yes	Yes	No	1.38	1.29-1.46	0.32(0.031)	0.26-0.38	Study1= 6.30e-11 Study 2= 0.026 Study3= 0.008 Study4=0.003	<0.01
	Model 4	Yes	No	Yes	1.38	1.29-1.46	0.32 (0.030)	0.26-0.38	----	<0.01
	Model 5	Yes	Yes	Yes	1.38	1.29-1.46	0.32 (0 .031)	0.26 - 0 .38	----	<0.01
One-step models	Model 6	No	No	Yes	1.41	1.33-1.49	0.34(0.029)	0.29 - 0.40	----	<0.01
	Model 7	No	Yes	Yes	1.42	1.33-1.51	0.35(0.032)	0.29-0 .41	0.013	<0.01
	Model 8	Yes	No	Yes	1.37	1.30-1.45	0.32(0.029)	0.26 -0 .37	----------	<0.01
	Model 9	Yes	Yes	Yes	1.38	1.29-1.46	0.32 (0.031)	0.26- 0.38	Study1= 2.32 e-09 Study2=0 .042 Study3= 8.44e-13 Study4= 0.001	<0.01
	Model 10	Yes	No	Yes	1.38	1.29-1.46	0.32(0.030)	0.26- .38	0.196	<0.01
	Model 11	Yes	Yes	Yes	1.39	1.31-1.48	0.33(0.031)	0.27- 0 .39	Study1= .003 Study2= .049 Study3= .011 Study4= 3.56e-09	<0.01

N.B. all of results in this Table have been estimated by using STATA software, and the stata modules to fit the two-step models are logit , xtmelogit and metan; for one- step models, the models are logit and xtmelogit [128], Model 4 and 5 are estimated by SAS, and the code was written by Richard Riley, for more details see Appendix C

One-step Models

Odds Ratio

As can be seen in Table 4.2, there is very little difference for the odds ratio between models 6 to 11. The pooled effect size is slightly higher in models 6 and 7 which ignoring clustering of patients within studies, with odds ratio increasing from 1.37 to 1.41 and 1.42 respectively. All of the one-step models produce p-vaue <0.01, which indicate that there is a significant association between age and mortality rate, i.e. age is a PF.

Standard Error and Confidence Interval for $\hat{\beta}$

There is very little difference for the confidence interval and standard error between models 6 to 11; however the standard error for model 7, 9 and 11 are slightly higher, as this model account for the residual variance for the same value(s) of the PF in each study.

Two-step versus One-step Models

Odds Ratio

There is very little difference for the odds ratio between models 2 to 11 and thus little difference between two-step and one-step; however one-step models 6, and 7 increase the pooled log odds ratio slightly from 0.32 to 0.34 and 0.35; which equates to an increase of the odds ratio from 1.38 to 1.41 and 1.42 respectively. This is due to model 6 and 7 ignoring the clustering of patients within studies. Also, it is clear that age is a PF as all of the one-step and two-step models yielded P-value < 0.01, and the confidence interval of the PF (e.g. age) for all models do not include one. This indicates that all of the one-step and two-step models have the same conclusion whether the meta-analysis models account for the correlation between parameters, taking into account the clustering of the patients across studies or consider the residual variation for the same values of the PF (e.g. age).

Standard Error and Confidence Interval for $\hat{\beta}$

There is very little difference for the confidence interval, and standard error for the pooled log odds ratio between models 2 to 11; the smallest standard errors seem to occur when clustering is ignored (model 6), do not account for the residual variation and account for correlation between $\hat{\alpha}_i$ and $\hat{\beta}_i$ (model 8). Also, the standard error for models 3, 7, 9, and 11 are slightly higher as these models account for the residual variance for the same value(s) of the PF in each study. However, one-step and two-step are giving very similar and often identical answers.

To sum up, there are only minor differences between the results of one-step and two-step models for this data, as all of the estimated values of pooled effect size and its standard error are approximately the same. It is difficult to decide the best model at this stage. Clearly, this is just one dataset, so a simulation study is needed to confirm these results (see chapter 5). Indeed, given that ignoring clustering and ignoring residual variation leads to lower standard errors, it is important that in other datasets this does not actually lead to overly-precise conclusions.

4.6.2 Application for Binary PF

For illustration, I now consider how models 2 to 11 perform when the factor of interest is binary. Table 4.3 gives the results of the IPD meta-analysis models when age, x_{ik}, is made binary (i.e. $age = 1\ if\ age \geq 40\ and\ age = 0\ if\ age < 40$). Note that the residual variation for the same values of the PF in each study can only be estimated when the PF is continuous, so these models (model 3, 5, 7, 9 and 11) are excluded. Thus, the focus is only on whether the IPD meta-analyses models account for the clustering of patients across studies and whether they account for the correlation between parameters $\hat{\propto}$ and $\hat{\beta}$. Note that the dichotomisation of age is merely for illustrative purpose.

Two–step Models

Odds Ratio

There is no difference in the pooled odds ratio estimate between models 2 and 4 as the pooled odds ratio is the same (2.45). Both models produce P-value <0.01, this again showed that age is still a PF.

Standard Error and Confidence Interval for $\widehat{\beta}$

There are barely any differences in the standard error and confidence interval between model 2 and 4.

One-step Models

Odds Ratio

Again, there is only slight difference for the odds ratio between models (Table 4.4). The odds ratio is slightly higher in model 6; this may be because it ignores the clustering of patients across studies.

Standard Error and Confidence Interval for $\widehat{\beta}$

There is very little difference in the confidence interval and standard error between models. However the standard error for model 6 is slightly smaller potentially, due to the lack of clustering.

Table 4. 3: Meta-analysis results from two-step and one-step meta-analysis to the TBI data; when age is binary (1 if age >= 40, 0 if age < 40).

Model type	Model number	Accounting for clustering	Accounting for residual variation	Accounting for correlation	Odds ratio	95% confidence interval for Odds ratio	Ln (odds ratio) and s.e	95% confidence interval for Ln Odds ratio	P-value for pooled effect-size, $\hat{\beta}$
Two-step Meta-analysis	Model 2	Yes	No	No	2.45	2.05 to 2.94	0.90 (0.092)	0.72 to 1.08	<0.01
	Model 4	Yes	No	Yes	2.45	2.05 to 2.94	0.90 (0.091)	0.72 to 1.08	<0.01
One-step meta-analysis	Model 6	No	No	Yes	2.59	2.18 to 3.10	0.95 (0.090)	0.78 to 1.13	<0.01
	Model 8	Yes	No	Yes	2.46	2.05 to 2.95	0.90 (0.091)	0.72 to 1.08	<0.01
	Model 10	Yes	No	Yes	2.48	2.08 to 2.95	0.91 (0.091)	0.73 to 1.08	<0.01

N.B. all of results in this Table have been estimated by using STATA software, and the stata modules to fit the two-step models are logit, xtmelogit and metan; for one- step models, the models are logit and xtmelogit [128]

Two-step versus One-step Models

Odds Ratio

There is a little difference between the odds ratio estimate for the five models; however, it is slightly higher in model 6 which ignores the clustering across studies.

Standard Error and Confidence Interval for $\widehat{\beta}$

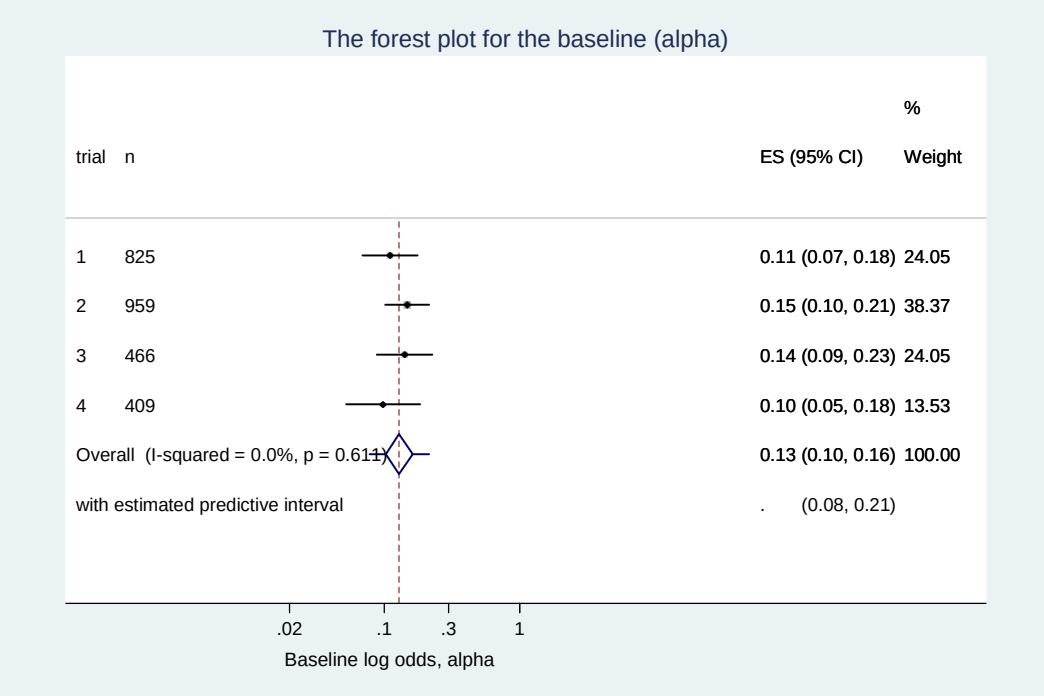
There is barely difference of the standard error and confidence interval among the 5 models. However, the standard error of model 6 is slightly smaller (0.090), this again potentially due to the lack of clustering (Table 4.3).

To sum up, whether the PF is continuous or binary in this case, generally, there is barely difference between the results for one-step and two-step models. So it is difficult to decide which approach is the best (one-step or two-step); and whether ignoring clustering or residual model variation really matters. I undertake a simulation study to investigate and examine some of these models in the next chapter.

Heterogeneity in $\propto$

Now, I assess the heterogeneity for alpha to examine whether there is a substantial variation of patients across studies for the TBI dataset. I found that there is no heterogeneity across studies for the baseline of the disease ($\propto$) as $I^2 = 0\%$ (see Figure 4.1). This might be the reason that there is no substantial difference between IPD meta-analysis models when they account for the clustering of patients across studies and when they ignore it (as there is no substantial variation across studies).

Figure 4.1: The forest plot for the baseline of the disease ($\propto$), in each IPD study for the TBI dataset



4.7 Discussion and Limitation

In this chapter I have proposed statistical models for meta-analysis binary outcome data in the presence of the one PF (age) for four IPD studies of TBI dataset. In particular, I show all possible two-step and one-step fixed-effect models for IPD studies that ignore or account for the clustering across studies and the residual variance for the patients who have the same value of the PF. In addition, I showed theoretically (see Appendix B) that correlation between parameters

(i.e. the intercept and the pooled effect size of PF 'age') does not effect on the pooled effect size of the binary PF when $\tau_\alpha^2 = 0$. However, the correlation may effect on the pooled effect size of the PF when $\tau_\alpha^2 \neq 0$, whether the PF is binary or continuous (see end of chapter 5). Application was made to 4 IPD studies of TBI examining the effect of age on six month mortality risk of death. The results show it is difficult to decide which model is the best at this stage because the results of pooled estimate of the PF, its standard error and confidence interval are all similar for most of the IPD models; for that reason, in chapter (5) a simulation study is made to decide which models are most accurate.

In Riley et al.[112] the authors developed IPD meta-analysis models for a binary outcome; in particular, they focused on the event-risk across studies and they considered residual variation. However, their focus was mainly on the event-risk across studies not on the effect size of the PF as undertaken in this chapter. In *Goldstein* et al.[129] the authors focused on the one-step meta-analysis by using multilevel models to consider the clustering across studies. However, they did not consider the two-step meta-analysis approach or even investigate whether there is difference in the results by following this approach.

In Turner et al.[66] the authors used multilevel model meta-analysis for a binary outcome, and they considered fixed and random effects. The authors stated that "*we explore the potential of multilevel models for meta-analysis of trials with binary outcomes for both summary data, such as log-odds ratios, and individual patient data. Conventional fixed effects and random effects models are put into a multilevel model framework, which provides maximum likelihood or restricted maximum likelihood estimation"*. They also stated that "*If random trial effects are used the covariance between these and the random treatment effects should be included; the resulting model is equivalent to a bivariate approach to meta-analysis*". However, again the

authors did not consider the two-step approach and whether there is a difference between one-step approach by using multilevel models and two-step approach, and which one is the best.
In Mathew et al.[130] the authors compared between one-step and two-step approaches. They found that both models are coincide, if they assumed the model with fixed treatment and fixed trial effects.

There are a few limitations in this chapter, one of these limitations is that the IPD meta-analysis models are fitted and developed for one variable, and I used fixed-effect method to estimate the pooled effect size of the PF (as there is no heterogeneity across studies in TBI dataset). However, this work should be extended in other datasets later by considering random-effects for the pooled effect size of the PF and multivariable analyses. Also, Cox proportional hazard model may be considered if the outcome is time until the event occurs, rather than logistic regression.

To sum up, by applying the 10 IPD meta-analysis models for TBI dataset, it yielded that there is no difference between one-step and two-step method. In particular, the IPD meta-analysis models yielded the same results whether they account for the clustering of patients within studies naturally in two-step models, or by including an indicator variable in one-step models. Also there is no difference between the results for IPD meta-analysis models if they account for the correlation between parameters naturally in one-step models, or by forcing the model to account for by using BFMA model in two-step models, or even if by including the residual variation for the same value of the PF that allows for unexplained variation due to the covariate, x_{ik} , to be accounted for. The similar results that yielded from the 10 IPD meta-analysis models could be because there is a little variation across studies, as there is no heterogeneity across TBI studies for$\propto$. In other datasets with more variation in $\propto_i$, my findings may change. Thus, a simulation study is undertaken in chapter 5 to examine whether there is generally a difference between one-step and two-step approach, and I aim to find an approach to recommend.

CHAPTER 5

A SIMULATION STUDY TO ASSESS WHETHER THE CLUSTERING OF PATIENTS WITHIN STUDIES CAN BE IGNORED IN IPD META-ANALYSIS OF A SINGLE PROGNOSTIC FACTOR

5.1 Introduction

Clustered data arise frequently in medical research (e.g. clustering of patients within different hospitals, or different general practices, or even individual doctors themselves). In an IMPF, the IPD are synthesised from different studies with potentially different study design, study populations and study locations. Thus, there is clustering of patients within studies, and this hierarchical structure may represent an important source of variation. The question thus arises; *'what is the impact of ignoring the clustering of patients within studies?'* Hogan et al,[131] states that *'Ignoring important sources of variation in any analysis can lead to incorrect confidence intervals and P values'*. However in chapter 2 I found that 6 of the 20 IMPF studies did not state whether they adjusted for clustering. Further in chapter 4 I analysed the TBI data and showed that conclusions and estimates were barely affected by whether clustering was accounted for or not. Although this was a single example shows that accounting for clustering may not always be necessary- but should this be the general advice?

In this chapter I perform a simulation study to assess three of the IPD meta-analysis logistic regression models that I introduced in chapter 4, in order to formally assess whether the clustering within studies can be ignored. These models are the two-step approach model 2, that accounts for the clustering of patients within studies by applying two-step; the one-step approach that ignores the clustering of patients within studies (model 6); and the one-step approach that accounts for the clustering of patients within studies by using indicator variable (model 8). I

undertake my simulation for a single binary PF, and then a single continuous PF. The simulation also allows a more formal comparison of the one-step versus two-step framework. Not that, for simplicity and limited time I will not consider the models that account either for the residual variations for the PF, or the correlation between parameters across studies.

5.2 Simulation for a Binary PF

In this section I describe in detail the simulation procedure used to generate IPD from multiple studies to fit the meta-analysis models to the data, and to evaluate their performance. Initially, just a binary PF is considered, and a binary outcome (e.g. death). I want to emphasise here that all the simulations I perform in this thesis generate data at the patient-level. Further, the initiative STATA code for these simulations was written by Boliang Guo according to my specifications and design.

5.2.1 The Simulation Procedure

The procedure used in my simulation can be broken down in 7 steps as follows:

Step 1: The number of studies in the meta-analyses is chosen as 5 and 10 respectively. Note that the number of studies was fixed in a simulation.

Step 2: randomly sample the number of patients in each study from a uniform distribution $n\sim U(a,b)$; $a=30, b=100$ for small sample size and $a=30$ and $b=1000$ for large sample size, a and b were fixed per simulation.

Step 3: For each patient in each trial, randomly sample a PF value using for a binary PF using a Bernoulli distribution, where $x_{ik} \sim Berrnoulli(prevelance)$. The prevalence was 0.5 or 0.2. For a continuous PF a normal distribution is used to generate the data, where $x_{ik} \sim N(\mu, \sigma^2)$ where$\mu = 4\ and\ \sigma = 1.5$. Note that i refers to the number of studies, and k refers to the number of patient.

Step 4: Sample the binary outcome ($Y_{ik}\ where\ Y_{ik} = 1\ for\ dead\ and\ Y_{ik} = 0\ \ for\ alive$) for each patient according to a specified relationship between the outcome and the PF. Where Y_{ik} $\sim Bernoulli\ (P_{ik})$, $\log\left(\frac{P_{ik}}{1-P_{ik}}\right) = \propto_i + \beta\ x_{ik}$, $\propto_i \sim N(\propto, \sigma_\alpha^2)$, where β is the chosen PF effect (log odds ratio) and $\propto_i$ could vary across the i studies according to a chosen $N(\propto, \sigma_\alpha^2)$ distribution.

Step 5: Fit each of the three different meta-analyses models to the data generated as follows:

Model 2 (Two-step method) recall in chapter 4 - in the first step, $\hat{\beta}_i$ is estimated for each IPD study separately and the pooled effect size is then estimated by using fixed-effect meta-analyses of the $\hat{\beta}_i$obtained:

$$First\ step: \qquad logit(P_{ik}) = \propto_i + \beta_i x_{ik}$$

$$Second\ step: \qquad \hat{\beta}_i = \beta + e_i$$

$$e_i \sim N\left(0, Var\ (\hat{\beta}_i)\right)$$

$$\hat{\beta} = \frac{\sum_{i=1}^{k} w_i \hat{\beta}_i}{\sum_{i=1}^{k} w_i}$$

Where $w_i = \frac{1}{s_i^2}$, refers to the relative weight for each study i.

Model 6 **(**One-step method ignoring clustering) in this model the pooled value for $\hat{\beta}$ is estimated by pooling all of the IPD data together without considering the clustering of patients within studies and the model is given as

$$logit(P_{ik}) = \propto + \beta x_{ik}$$

Model 8 **(**One-step method accounting for clustering) in this model the pooled value for $\hat{\beta}$ is estimated in one step by accounting for the clustering of patients within studies by including study indicators, that allow a separate intercept term for each study, and the model is given as

$$logit(P_{ik}) = \propto_i + \beta x_{ik}$$

Step 6: Repeat steps 1 to 5 a thousand times, keeping the chosen range of sample sizes, number of studies and parameter values as before in each step. This resulted in 1000 $\hat{\beta}$ values and 1000 standard error of $\hat{\beta}$ values for each model.

Step 7: Repeat step 1 to 6 for a different set of simulation criteria; that is choose again the number of studies, sample size distribution, x_{ij} distribution and values, and the true α_i, β, and σ_α^2 values in steps one to four, then fit the models to the generated data, and repeat 1000 times.

For a binary PF, Table 5.1 shows the different permutations I chose according to different α_i, β, σ_α^2 and prevelance of $x_{ik} = 1$ values. Note that true $\propto$ here means the average log odds of the event (e.g. death) when the PF is zero, I chose this value according to the TBI data; β means the true pooled effect size of the PF and again I chose the value of this parameter equals 0.90 according to the TBI data then consider other situations by assuming $\beta = 0 \ or \ 0.10$, relating to no prognostic effect and small effect respectively. Also I chose $\propto = -1.27$ again according the TBI dataset and I allow different standard deviations (SD) for $\propto$ (i.e. SD=0 such that baseline risk is the same across studies; and 'SD= 0.25 or 1.5' which assumes a small and slightly larger baseline risk respectively). Note that when SD=1.5 of the baseline risk $\propto$, 95% range for the baseline $log\left(\frac{p}{1-p}\right)$ across studies is between (-1.27$\pm$ 2×1.5), and the rearranging gives a 95% range for the baseline probability from 0.014 and 0.849. So, there is a substantial variation for the baseline risk $\propto$ across studies. The standard deviation for $\propto$ here indicates clinically that the

variation in baseline risk, $\propto$ across studies might come from other factors such as the treatments being used in study, its location, measurement techniques etc.

Table 5.1: The possible values considered for $\propto$, standard deviation of $\propto$, $\boldsymbol{\beta}$ and prevalence of the binary PF in the simulation scenarios.

Scenario no.	*True* $\propto$	*Standard deviation for* $\propto$	*True* β	*Prevalence of the true PF*
1	-1.27	0	0.90	0.5
2	-1.27	0	0.10	0.5
3	-1.27	0	0.00	0.5
4	-1.27	0.25	0.90	0.5
5	-1.27	0.25	0.10	0.5
6	-1.27	0.25	0.00	0.5
7	-1.27	0	0.90	0.2
8	-1.27	0	0.10	0.2
9	-1.27	0	0.00	0.2
10	-1.27	0.25	0.90	0.2
11	-1.27	0.25	0.10	0.2
12	-1.27	0.25	0.00	0.2
13	-1.27	1.5	0.90	0.2
14	-1.27	1.5	0.10	0.2
15	-1.27	1.5	0.00	0.2
16	-1.27	1.5	0.90	0.5
17	-1.27	1.5	0.10	0.5
18	-1.27	1.5	0.00	0.5

Each scenario was considered for 5 or 10 studies, and 30 to 100 (for small sample size) or 30 to 1000 (for large sample size) patients.

5.2.2 Evaluating the Performance of Statistical Models

For each simulated scenario, 1000 $\hat{\beta}$ and their standard error were available for each model. The assessment of the three models was examined in terms of different criteria such as bias, mean square error (MSE), and coverage for all parameters estimates and their standard errors for all scenarios.

Assessment of Bias

Bias is the deviation of the average of the pooled effect size estimates from the true value ($\bar{\hat{\beta}} - \beta$); the smaller the bias, the better the estimation method and on unbiased method would have zero bias[132-134]. Percentage bias is another statistic to summarise bias; it is given as $\left(\frac{\bar{\hat{\beta}} - \beta}{\beta}\right) * 100$. Larger percentage bias indicates a worse estimation and if the percentage bias is greater than 10% , this indicates that the bias is meaningful[134]. Note that with 95% confidence interval and 1000 simulations studies, it yielded a small difference between the true bias and the estimated bias by using the following equation

$$\delta = \frac{(Z_{1-(\alpha/2)}\ \sigma)^2}{B}$$

Where B is the number of simulation studies, $Z_{1-(\propto/2)}$ is the $1 - \propto/2$ quantile of the standard normal distribution, σ^2 is the variance of the pooled effect size of the PF, and δ is the specified level of accuracy of the pooled effect size of the PF (i.e. the accepted difference between the true values β).

Assessment of Accuracy

Mean square error (MSE)[135] is used to assess the overall accuracy of the statistical model and it is given as

$$MSE = (\bar{\hat{\beta}} - \beta)^2 + (SE(\hat{\beta}))^2$$

Where $(\bar{\hat{\beta}} - \beta)^2$ refers to bias of the difference between the average pooled estimate, $\bar{\hat{\beta}}$, and the true pooled effect size, β, for the 1000 estimates and $(SE(\hat{\beta}))^2$ refers to the mean standard error of $\hat{\beta}$ in one-step and two step meta-analyses. It is considered as a useful method for the assessment of estimation accuracy as it includes both of the bias and the variability.

Assessment of Coverage

Burton et al.[132] state that '*The coverage of a confidence interval is the proportion of times that the obtained confidence interval contains the true specified parameter value'*. The nominal coverage rate in this chapter is 95%, and a good estimation method will give an observed coverage in the simulation the same as the nominal coverage rate. If the observed coverage rate is above the nominal average rate, then the estimation method is too conservative, this means that too few samples will find significant results when the true effect is non-zero (i.e. the factor is prognostic); thus this leads to a loss of the statistical power. If the coverage rate is lower than the nominal coverage rate; this leads to over-confidence and too many samples then will yield significant results when the true effect is zero (i.e. the factor is not prognostic); this leads to increased type I error and is anti-conservative. To check whether the coverage is suitable in the 1000 simulations in each scenario, I expected coverage percentage should lie between $p \pm 1.96\, s.e\,(p)$ where $s.e\,(p) = \sqrt{p(1-p)/1000}$, and $p = 0.95$ assuming that the true coverage is 95%. The $s.e(p)$ in my simulations equals 0.006892, so the observed coverage should lie between 0.936 and 0.964. If the estimated coverage lies outside this interval, this indicates a problem.

5.3 Simulation Results for the Binary PF

In this section I describe the simulation results when x_{ik} is a binary PF in relation to the three IPD meta-analysis models for all of the 18 scenarios (Table 5.1) given studies with a small sample size (30 to 100) or large sample size (30 to 1000) , for a meta-anlysis of 5 studies or 10 studies or the prevelance of the PF is either 0.5 and 0.2.

5.3.1 Simulation Results for Small Sample Size (30 to 100) and n=5

Table 5.2 gives the simulation results for the three models when the prevalence=0.5 and standard deviation of $\propto$ =0 and 0.25 respectively for the first 6 scenarios (1 to 6). For all of these 6

scenarios, all models give a bias almost zero, and the percentage bias is almost 6%. MSE and the mean of the standard error for the three models is approximately the same, and their coverage is always close to 95%.

I changed the prevalence of x_{ik} to be 0.2 and repeated the above simulations; the results are shown in Table 5.3 (scenarios 7 to 12). In the first three scenarios (7, 8, and 9), the bias and percentage bias was again small for all models, and the MSE was approximately the same; however, the two-step models gave the highest percentage bias and in scenario 9 its coverage was slightly too large (and does not fall between 0.936 and 0.964), whilst the other models had coverage closer to 95%. In the other three scenarios (10, 11 and 12) with standard deviation of $\propto = 0.25$; the bias is still small for both one-step models (0.03 or less), although their percentage bias is slightly above 10 % in scenario 11 where the true pooled value is 0.1. More concerning, the bias and percentage biases are large for two-step method. For example, in scenario 11 there is an upward bias of 0.11 and a percentage bias of 109%.

Finally, I changed the standard deviation of $\propto$ to be 1.5 and repeated the simulation when the prevalence is 0.2 and 0.5 respectively. The results shows that in scenario 13 (where $\beta = 0.9$) , model 6 (ignoring clustering) has a large downward bias of –0.21 with a percentage bias of -23.5% and a low coverage of 87.6% (which does not fall between 0.936 and 0.964), reflecting a small mean standard error. Its percentage bias is similar in scenario 14. Clearly, when the standard error of $\propto$ is large, the one-step ignoring clustering (and thus assuming a fixed $\propto$ across studies) is inappropriate. The two-step method (in scenarios 13 to 15) continues to have a coverage somewhat greater than 95%, and in scenario 14 its upward bias is upward large again. In contrast the one-step accounting for clustering is approximately unbiased, and has suitable coverage. Also, when the prevalence is 0.5 (in scenarios 16 to 18), the two-step method yield no bias however it is still has a coverage somewhat greater than 95% because the estimated

confidence interval again falls between 0.936 and 0.964. The one-step method ignoring clustering still yields inaccurate results (downward bias and low coverage), see Table 5.4.

Figure 5.1 shows the pooled effect size of the PF and its standard error for the 1000 observations for the two-step model, one-step model accounting for the clustering and one-step model ignores the clustering of patients across studies, when the SD $(\propto) = 0.25\ and$ 1.5, and the true β = 0.1 and 0.9 respectively, for scenario 11 and 13 in Table 5.4. In particular, Figure 5.1a gives the scatter plot for the pooled effect size of the two-step model versus one-step accounting for the clustering for 1000 observations, when the SD $(\propto) = 0.25$, and the true β = 0.1. The estimated pooled effects size of the PF from two-step model generally lies above the line of equality, which reflect that the pooled effect size of PF in the two-step model is somewhat overestimated (large bias) compared to one-step model accounts for clustering. Also, Figure 5.1b shows the standard errors for the same models, again the estimated values of the standard error in two-step model are generally slight above the line of equality. This again indicates that the standard error is slightly higher in two-step model compared to one-step model, if there is variation in the baseline risk $(\propto)$.

Figure 5.1c gives the scatter plot for the pooled effect size of the one-step model ignores the clustering versus one-step model accounts for the clustering for 1000 observations, when the SD $(\propto)$=1.5 , and the true $\beta = 0.9$. The estimated pooled effects size of the PF from one-step ignores clustering model lie below the line of equality, which reflect that the pooled effect size of PF in one-step ignoring clustering model is somewhat underestimated (downward bias) compared to one-step model accounts for clustering. Also, Figure 5.1d shows the standard errors for the same models, again the estimated values of the standard errors in one-step ignores clustering model are slightly under the line of equality. This again indicates that the standard

error is slightly smaller in one-step ignores clustering model compared to one-step accounts for clustering model, if the variation for the baseline risk $\propto = 1.5$.

Figure 5.1: The scatter plots for the pooled effect size and its standard error for two-step models, one-step model that accounts for clustering and the one-step model that ignore the clustering, in scenario 11 and 13 in Table 5.4

Figure 5.1a: The scatter plot for the pooled effect-size of the two-step versus one-step model accounts for clustering, for scenario 11 with n=5, SD $(\propto) = 0.25, \beta = 0.1$ and the sample size between 30 to 100 for 1000 simulation study.

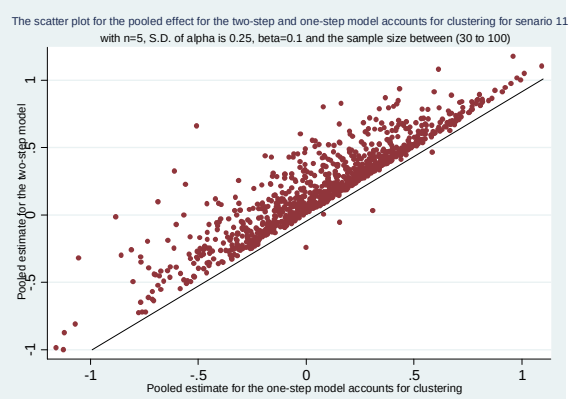


Figure 5.1b: The scatter plot for the standard error for the two-step versus one-step model accounts for clustering, for scenario 11 with n=5, SD $(\propto) = 0.25, \beta = 0.1$ and the sample size between 30 to 100 for 1000 simulation study.

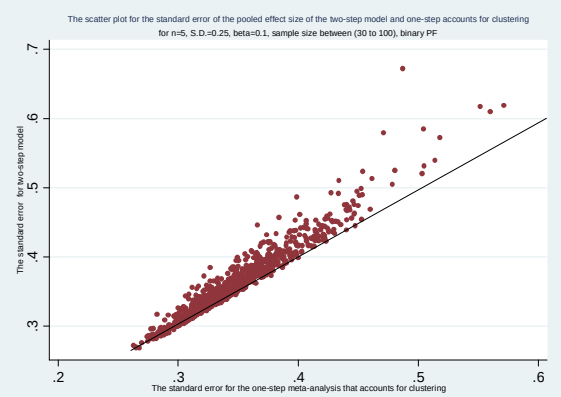


Figure 5.1c: The scatter plot for the pooled effect-size of the one-step accounts for clustering versus one-step model ignore clustering, for scenario 13 with n=5, SD $(\propto) = 1.5, \beta = 0.9$, and the sample size between 30 to 100, for 1000 simulation study.

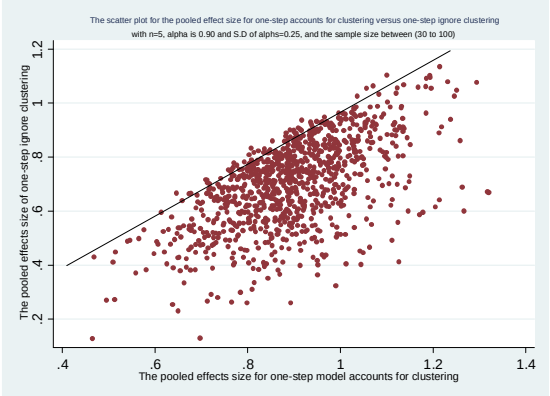


Figure 5.1d: The scatter plot for the standard error for the one-step accounts for clustering versus one-step model ignore clustering, for scenario 13 with n=5, SD$(\propto) = 1.5$, $\beta = 0.9$, and the sample size between 30 to 100, for 1000 simulation study.

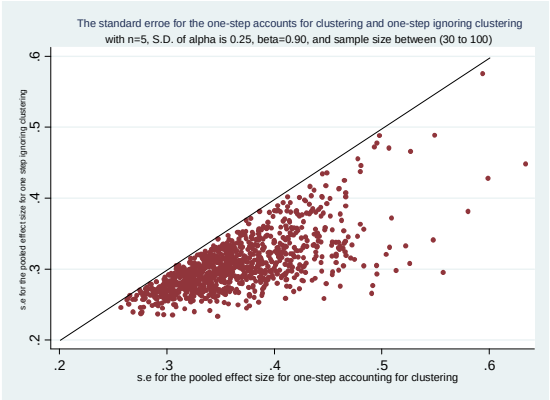


Table 5. 2: Simulation results for the pooled effect size,$\widehat{\boldsymbol{\beta}}$, for three models of IPD meta-analysis, two-step, one-step ignoring clustering and one-step including indicator variable, with prevalence =0.5and the sample size with-in each study is between 30 to 100 observations, with the number of studies n=5 and 1000 simulations, the true values of the pooled effect size,$\boldsymbol{\beta}$, is shown in the table; the standard error for $\propto$ is 0 and 0.25.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (SD of $\propto$)	*prevalence*	$\boldsymbol{\beta}$	Mean of $\widehat{\boldsymbol{\beta}}$	Bias	percentage bias*	MSE of $\boldsymbol{\beta}$	Percentage coverage	Mean s.e. of $\widehat{\boldsymbol{\beta}}$
1	5	Model 2	Two-step	-1.27(0)	0.5	0.9	0.91	0.01	0.99	0.03	94.90	0.17
		Model 6	Ignoring clustering	-1.27(0)	0.5	0.9	0.91	0.01	1.25	0.03	94.90	0.16
		Model 8	One-step including indicator variable.	-1.27(0)	0.5	0.9	0.92	0.02	1.83	0.03	94.70	0.16
2	5	Model 2	Two-step	-1.27(0)	0.5	0.10	0.10	0.00	2.67	0.00	95.60	0.16
		Model 6	Ignoring clustering	-1.27(0)	0.5	0.10	0.10	0.00	2.37	0.00	95.60	0.16
		Model 8	One-step including indicator variable.	-1.27(0)	0.5	0.10	0.10	0.00	3.36	0.00	95.60	0.16
3	5	Model 2	Two-step	-1.27(0)	0.5	0	0.00	0.00	NA	0.02	95.00	0.16
		Model 6	Ignoring clustering	-1.27(0)	0.5	0	0.00	0.00	NA	0.02	94.90	0.16
		Model 8	One-step including indicator variable.	-1.27(0)	0.5	0	0.00	0.00	NA	0.02	94.90	0.16
4	5	Model 2	Two-step	-1.27(0.25)	0.5	0.9	0.90	0.00	0.35	0.03	95.80	0.17
		Model 6	Ignoring clustering	-1.27(0.25)	0.5	0.9	0.90	0.00	0.00	0.03	95.20	0.17
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.5	0.9	0.91	0.01	1.58	0.03	95.30	0.18
5	5	Model 2	Two-step	-1.27(0.25)	0.5	0.10	0.09	-0.01	-5.61	0.04	95.10	0.19
		Model 6	Ignoring clustering	-1.27(0.25)	0.5	0.10	0.09	-0.01	-5.80	0.04	94.60	0.18
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.5	0.10	0.10	0.00	-3.58	0.04	94.80	0.19
6	5	Model 2	Two-step	-1.27(0.25)	0.5	0	0.01	0.01	NA	0.04	95.50	0.19
		Model 6	Ignoring clustering	-1.27(0.25)	0.5	0	0.01	0.01	NA	0.04	94.70	0.19
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.5	0	0.01	0.01	NA	0.04	94.50	0.19

*Note that all the results are shown to two decimal places but the calculations such as bias, percentage bias, ... etc are based on full results (more than two decimal places). Hint: in the first two rows are different even though the bias are identical for the decimal places/

Table 5. 3: Simulation results for the pooled effect size, $\widehat{\boldsymbol{\beta}}$,, for three models of IPD meta-analyses (two-step, one-step ignoring clustering and one-step including indicator variable, with prevalence =0.2 and the sample size with-in each study is between 30 to 100 observations, with the number of studies n=5 and 1000 simulations, the true values of the pooled effect size,$\boldsymbol{\beta}$, is shown in the table; the standard error for $\propto$ is 0 and 0.25.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (SD of $\propto$)	Prevalence	True $\boldsymbol{\beta}$	Mean of $\widehat{\boldsymbol{\beta}}$	Bias	percentage bias*	MSE of $\boldsymbol{\beta}$	Percentage coverage	Mean s.e. of $\widehat{\boldsymbol{\beta}}$
7	5	Model 2	Two-step	-1.27(0)	0.2	0.9	0.83	-0.07	-8.13	0.10	95.80	0.32
		Model 6	Ignoring clustering	-1.27(0)	0.2	0.9	0.90	0.00	0.02	0.10	94.90	0.31
		Model 8	One-step including indicator variable.	-1.27(0)	0.2	0.9	0.91	0.01	1.13	0.10	94.70	0.31
8	5	Model 2	Two-step	-1.27(0)	0.2	0.10	0.10	0.00	-4.25	0.08	96.40	0.29
		Model 6	Ignoring clustering	-1.27(0)	0.2	0.10	0.10	0.00	-1.71	0.08	95.30	0.28
		Model 8	One-step including indicator variable.	-1.27(0)	0.2	0.10	0.10	0.00	0.40	0.09	95.40	0.29
9	5	Model 2	Two-step	-1.27(0)	0.2	0	0.01	0.01	NA	0.07	97.00	0.30
		Model 6	Ignoring clustering	-1.27(0)	0.2	0	0.01	0.01	NA	0.08	95.80	0.28
		Model 8	One-step including indicator variable.	-1.27(0)	0.2	0	0.01	0.01	NA	0.08	95.80	0.29
10	5	Model 2	Two-step	-1.27(0.25)	0.2	0.9	0.94	0.04	4.50	0.09	96.60	0.32
		Model 6	Ignoring clustering	-1.27(0.25)	0.2	0.9	0.90	0.00	-0.43	0.09	95.10	0.30
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.2	0.9	0.92	0.02	2.06	0.10	95.40	0.31
11	5	Model 2	Two-step	-1.27(0.25)	0.2	0.1	0.21	0.11	109.64	0.12	95.60	0.36
		Model 6	Ignoring clustering	-1.27(0.25)	0.2	0.1	0.08	-0.02	-16.18	0.12	95.50	0.34
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.2	0.1	0.09	-0.01	-14.91	0.12	94.90	0.35
12	5	Model 2	Two-step	-1.27(0.25)	0.2	0	0.11	0.11	NA	0.13	95.20	0.37
		Model 6	Ignoring clustering	-1.27(0.25)	0.2	0	-0.03	-0.03	NA	0.12	95.20	0.35
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.2	0	-0.03	-0.03	NA	0.13	95.10	0.35

* Note that all the results are shown to two decimal places but the calculations such as bias, percentage bias, ... etc are based on full results (more than two decimal places).

Table 5.4: Simulation results for the pooled effect size, β , for three models of IPD meta-analyses (two-step, one-step ignoring clustering and one-step including indicator variable), with prevalence =0.2 and 0.5 and the sample size with-in each study is between 30 to 100 observations, with the number of studies n=5 and 1000 simulations, the standard error for $\propto$ is 1.5.

Scenarios	No. of studies (n)	Model	Meta-analysis model	$\propto$ (SD of $\propto$)	Prevalence	True β	Mean of $\widehat{\beta}$	Bias	percentage bias*	MSE of β	Percentage coverage	Mean s.e. of $\widehat{\beta}$
13	5	Model 2	Two-step	-1.27(1.5)	0.2	0.9	0.91	0.01	0.87	0.13	96.20	0.38
		Model 6	Ignoring clustering	-1.27(1.5)	0.2	0.9	0.69	-0.21	-23.50	0.15	87.60	0.31
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.2	0.9	0.92	0.02	2.05	0.14	94.80	0.36
14	5	Model 2	Two-step	-1.27(1.5)	0.2	0.1	0.20	0.10	98.37	0.16	95.80	0.41
		Model 6	Ignoring clustering	-1.27(1.5)	0.2	0.1	0.07	-0.03	-27.36	0.12	95.70	0.33
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.2	0.1	0.10	0.00	-3.99	0.17	94.20	0.38
15	5	Model 2	Two-step	-1.27(1.5)	0.2	0	0.09	0.09	NA	0.28	96.89	0.41
		Model 6	Ignoring clustering	-1.27(1.5)	0.2	0	-0.02	-0.02	NA	0.22	94.00	0.33
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.2	0	0.00	0.00	NA	0.26	94.00	0.38
16	5	Model 2	Two-step	-1.27(1.5)	0.5	0.9	0.89	0.01	-1.11	0.01	95.40	0.11
		Model 6	Ignoring clustering	-1.27(1.5)	0.5	0.9	0.70	-0.20	-22.22	0.04	46.20	0.09
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.5	0.9	0.90	0.00	0.00	0.05	94.90	0.11
17	5	Model 2	Two-step	-1.27(1.5)	0.5	0.1	0.10	0.00	0.00	0.01	95.00	0.11
		Model 6	Ignoring clustering	-1.27(1.5)	0.5	0.1	0.08	-0.02	-20.00	0.04	93.90	0.09
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.5	0.1	0.10	0.00	0.00	0.05	94.80	0.11
18	5	Model 2	Two-step	-1.27(1.5)	0.5	0	0.00	0.00	NA	0.01	95.10	0.11
		Model 6	Ignoring clustering	-1.27(1.5)	0.5	0	0.00	0.00	NA	0.04	94.90	0.09
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.5	0	0.00	0.00	NA	0.05	94.70	0..11

* Note that all the results are shown to two decimal places but the calculations such as bias, percentage bias, ... etc are based on full results (more than two decimal places).

5.3.2 Simulation Results for Large Sample Size (30 to 1000) and n=5

Table 5.6 to 5.8 gives simulation results when the sample size is from 30 to 1000, for scenarios 1 to 18. The results are similar to the simulation results for small sample size (30 to 100) as just discussed. In particular, model 8 (one-step accounting for clustering) performs consistently well with little bias and adequate coverage as it nearly always falls between 0.936 and 0.964. However, in Table 5.7 model 6 (one-step ignoring clustering) again has a low coverage when β is large (55.8%), which does not fall between 0.936 and 0.964, and a significant downward bias of -0.2 (scenario 13). Further, the two-step model performs better here than when the sample size was 30 to 100. Indeed, there is very little bias anymore and the coverage is adequate except in scenario 8 for two-step and one-step methods as the estimated confidence interval does not fall between 0.936 and 0.964. Thus, the two-step and one-step accounting for clustering methods appear very similar here.

Table 5. 5 Simulation results for the pooled effect size, $\widehat{\boldsymbol{\beta}}$, for the three modelsof IPD meta-analysis,(two-step, one-step ignoring clustering and one-step including indicator varable), with prevelance= 0.50 and the sample size with-in each study is between 30 to 1000 observations, with the number of studies n=5 and 1000 simulation, the true values of the pooled effect size,$\boldsymbol{\beta}$, is shown on the table, the standard error for $\propto$ is 0 and 0.25

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (SD of $\propto$)	*prevalence*	β	Mean of $\widehat{\beta}$	Bias	percentage bias*	MSE of β	Percentage coverage	Mean s.e. of $\widehat{\beta}$
1	5	Model 2	Two-step	-1.27(0)	0.5	0.9	0.90	0.00	-0.36	0.01	95.10	0.09
		Model 6	Ignoring clustering	-1.27(0)	0.5	0.9	0.90	0.00	-0.32	0.01	95.20	0.09
		Model 8	One-step including indicator variable.	-1.27(0)	0.5	0.9	0.90	0.00	-0.14	0.01	95.40	0.09
2	5	Model 2	Two-step	-1.27(0)	0.5	0.10	0.1	-0.01	-6.87	0.01	94.00	0.08
		Model 6	Ignoring clustering	-1.27(0)	0.5	0.10	0.1	-0.01	-6.74	0.01	94.00	0.08
		Model 8	One-step including indicator variable.	-1.27(0)	0.5	0.10	0.1	-0.01	-6.69	0.01	94.00	0.08
3	5	Model 2	Two-step	-1.27(0)	0.5	0	0	0.00	NA	0.01	94.90	0.08
		Model 6	Ignoring clustering	-1.27(0)	0.5	0	0	0.00	NA	0.01	94.90	0.08
		Model 8	One-step including indicator variable.	-1.27(0)	0.5	0	0	0.00	NA	0.01	94.90	0.08
4	5	Model 2	Two-step	-1.27(0.25)	0.5	0.9	0.90	0.00	0.05	0.01	95.20	0.09
		Model 6	Ignoring clustering	-1.27(0.25)	0.5	0.9	0.89	-0.01	-0.67	0.01	94.50	0.09
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.5	0.9	0.90	0.00	0.38	0.01	95.00	0.09
5	5	Model 2	Two-step	-1.27(0.25)	0.5	0.10	0.1	0.00	-1.84	0.01	95.10	0.10
		Model 6	Ignoring clustering	-1.27(0.25)	0.5	0.10	0.1	0.00	-1.86	0.01	95.00	0.10
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.5	0.10	0.1	0.00	-1.38	0.01	95.00	0.10
6	5	Model 2	Two-step	-1.27(0.25)	0.5	0	-0.01	-0.01	NA	0.01	95.60	0.10
		Model 6	Ignoring clustering	-1.27(0.25)	0.5	0	-0.01	-0.01	NA	0.01	96.00	0.10
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.5	0	-0.01	-0.01	NA	0.01	95.60	0.10

* Note that all the results are shown to two decimal places but the calculations such as bias, percentage bias, ... etc are based on full results (more than two decimal places).

Table 5. 6 Simulation results for the pooled effect size, $\hat{\beta}$, for three models of IPD meta-analyses (two-step, one-step ignoring clustering and one-step including indicator variable), with prevalence=0.2 and the sample size with-in each study is between 30 to 1000 observations, with the number of studies n=5 and 1000 simulations, the true values of the pooled effect size,β, is shown in the table; the standard error for $\propto$ is 0 and 0.25.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (SD of $\propto$)	Prevalence	True β	Mean of $\widehat{\beta}$	Bias	percentage bias*	MSE of β	Percentage coverage	Mean s.e. of $\widehat{\beta}$
7	5	Model 2	Two-step	-1.27(0)	0.2	0.9	0.90	0.00	-0.46	0.01	95.80	0.11
		Model 6	Ignoring clustering	-1.27(0)	0.2	0.9	0.90	0.00	0.21	0.01	95.70	0.11
		Model 8	One-step including indicator variable.	-1.27(0)	0.2	0.9	0.90	0.00	0.39	0.01	95.90	0.11
8	5	Model 2	Two-step	-1.27(0)	0.2	0.1	0.10	0.00	2.14	0.01	96.80	0.10
		Model 6	Ignoring clustering	-1.27(0)	0.2	0.1	0.10	0.00	2.71	0.01	96.60	0.10
		Model 8	One-step including indicator variable.	-1.27(0)	0.2	0.1	0.10	0.00	2.84	0.01	96.70	0.10
9	5	Model 2	Two-step	-1.27(0)	0.2	0	0.00	0.00	NA	0.01	94.80	0.10
		Model 6	Ignoring clustering	-1.27(0)	0.2	0	0.00	0.00	NA	0.01	94.60	0.10
		Model 8	One-step including indicator variable.	-1.27(0)	0.2	0	0.00	0.00	NA	0.01	94.60	0.10
10	5	Model 2	Two-step	-1.27(0.25)	0.2	0.9	0.90	0.00	0.35	0.01	96.00	0.11
		Model 6	Ignoring clustering	-1.27(0.25)	0.2	0.9	0.89	0.01	-0.91	0.01	95.60	0.11
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.2	0.9	0.90	0.00	0.16	0.01	95.80	0.11
11	5	Model 2	Two-step	-1.27(0.25)	0.2	0.1	0.11	0.01	5.65	0.02	94.40	0.12
		Model 6	Ignoring clustering	-1.27(0.25)	0.2	0.1	0.10	0.00	-5.13	0.02	94.80	0.12
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.2	0.1	0.10	0.00	-4.32	0.02	94.50	0.12
10	5	Model 2	Two-step	-1.27(0.25)	0.2	0	0.01	0.01	NA	0.02	94.80	0.12
		Model 6	Ignoring clustering	-1.27(0.25)	0.2	0	0.00	0.00	NA	0.02	95.40	0.12
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.2	0	0.00	0.00	NA	0.02	94.80	0.12

* Note that all the results are shown to two decimal places but the calculations such as bias, percentage bias, ... etc are based on full results (more than two decimal places).

Table 5. 7: Simulation results for the pooled effect size, β , for three models of IPD meta-analyses (two-step, one-step ignoring clustering and one-step including indicator variable), with prevalence =0.2 and the sample size with-in each study is between 30 to 1000 observations, with the number of studies n=5 and 1000 simulations, the standard error for $\propto$ is 1.5.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (SD of $\propto$)	Prevalence	True β	Mean of $\widehat{\beta}$	Bias	percentage bias*	MSE of β	Percentage coverage	Mean s.e. of $\widehat{\beta}$
13	5	Model 2	Two-step	-1.27(1.5)	0.2	0.9	0.90	0.00	-0.43	0.02	95.50	0.13
		Model 6	Ignoring clustering	-1.27(1.5)	0.2	0.9	0.70	-0.20	-22.19	0.07	55.80	0.11
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.2	0.9	0.90	0.00	-0.38	0.02	95.30	0.13
14	5	Model 2	Two-step	-1.27(1.5)	0.2	0.1	0.10	0.00	2.98	0.02	95.20	0.14
		Model 6	Ignoring clustering	-1.27(1.5)	0.2	0.1	0.07	-0.03	-28.96	0.02	94.30	0.12
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.2	0.1	0.10	-0.01	-8.08	0.02	94.90	0.14
15	5	Model 2	Two-step	-1.27(1.5)	0.2	0	0.01	0.01	NA	0.02	95.70	0.14
		Model 6	Ignoring clustering	-1.27(1.5)	0.2	0	0.00	0.00	NA	0.02	95.20	0.12
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.2	0	0.00	0.00	NA	0.02	95.80	0.14
16	5	Model 2	Two-step	-1.27(1.5)	0.5	0.9	0.86	-0.04	-4.44	0.09	96.00	0.31
		Model 6	Ignoring clustering	-1.27(1.5)	0.5	0.9	0.67	-0.23	-25.56	0.13	82.80	0.25
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.5	0.9	0.91	0.01	1.11	0.10	95.00	0.30
17	5	Model 2	Two-step	-1.27(1.5)	0.5	0.1	0.10	0.00	0.00	0.09	96.40	0.31
		Model 6	Ignoring clustering	-1.27(1.5)	0.5	0.1	0.07	-0.03	-30.00	0.07	94.80	0.26
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.5	0.1	0.10	0.00	0.00	0.10	94.70	0.30
18	5	Model 2	Two-step	-1.27(1.5)	0.5	0	-0.01	-0.01	NA	0.09	96.5	0.22
		Model 6	Ignoring clustering	-1.27(1.5)	0.5	0	0.00	0.00	NA	0.07	95.7	0.18
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.5	0	-0.01	-0.01	NA	0.10	95.2	0.21

*** Note that all the results are shown to two decimal places but the calculations such as bias, percentage bias, … etc are based on full results (more than two decimal places).**

5.3.3 Simulation Results for n=10 Studies

I repeated scenarios 1 to 18 for n=10 studies, and considered 30 to 100 and 30 to 1000 sample size . The results are shown in Appendix D. Conclusions are generally the same as before. The one-step accounting for clustering performs consistently well, with little bias and coverage close to 95%. The two-step is comparable to this method, and the upward bias previously seen in the n=5 simulations for scenario 11, and for 30 to 100 patients is almost gone. However, the one-step ignoring clustering continues to perform poorly when standard deviation of $\propto$ is large, with downward bias in the PF effect and too low coverage and too small standard error when $\beta = 0.9$.

5.4 Simulation Results for a Continuous PF

I repeated step 1 to 7 of the simulation procedure described in section 5.2.1, but now for a continuous PF sampled for each patient from a normal distribution $N\sim(\mu, \sigma^2)$. I assumed that $\mu = 4$, and standard error =1.5, this is similar to the values that I found in TBI data when looking at the distribution of age per 10 years.

Six different scenarios were considered as shown in Table 5.8. β now relates to the effect of a one-unit increase on the log-odds of death, and is specifying a true linear trend between x_{ik} and $\log\left(\frac{P_{ik}}{1-P_{ik}}\right)$. The scenario used $\propto = -2.10$, and $\beta = 0.3$, which is again similar to the estimated values from the TBI data, and also a small value of β, say of 0.1 and no effect of 0. The standard error of $\propto$ was either small (0.2) or large (1.5). The scenarios were considered for 5 and 10 studies, and 30 to 100 or 30 to 1000 patients.

Table 5. 8 The possible scenarios of $\propto$ and $\boldsymbol{\beta}$ for the continuous PF

Scenario no.	*True* $\propto$	*SD for* $\propto$	*True* β
1	-2.10	0.2	0.3
2	-2.10	0.2	0.1
3	-2.10	0.2	0
4	-2.10	1.5	0.3
5	-2.10	1.5	0.1
6	-2.10	1.5	0

5.4.1 Simulation Results for the Number of Studies n=5.

Simulation Results for Small Sample Size (30 to 100)

Table 5.9 gives the simulation results for scenario 1 to 6 for each of model 2, 6 and 8, for n=5 studies and 30 to 100 patients per study. For the first three scenarios, when the standard error of $\propto = 0.20$, there is a little difference among the three models. In particular, all models give little bias and a percentage bias less than 10 %. MSE is approximately the same in all models and the coverage is approximately the same, generally slightly over 95%. For scenarios 4 and 5 when standard deviation of $\propto = 1.5$, there is a downward bias for model 6 (ignoring clustering) and the percentage bias is greater than 10%. The coverage is much lower than 95% in model 6 when β is large (scenario 5). In scenarios 5 and 6 the coverage is too large for the two-step model (greater than 97%), whilst the one-step accounting for clustering (model 8) performs best with little bias and suitable coverage.

Simulation Results for Large Sample Size (30 to 1000)

Table 5.10 gives the simulation results for scenarios 1 to 6, for n=5 studies and now a sample size of 30 to 1000. The findings are as above, but the two-step now performs adequately in all scenarios; i.e. the coverage is much closer to 95% (as it falls between 0.936 and 0.964), even in scenario 4 to 6. The one-step accounting for clustering is consistently good again, but the one-step ignoring clustering still performs poorly in scenarios 4 and 5 (i.e. when the standard deviation of $\propto$ is large and there is a true PF effects > 0).

5.4.2 Simulation Results for the Number of Studies n=10

The simulation results for n=10 studies are shown in Appendix D. The results are the same as discussed for n=5, whether the sample size for each IPD study lies between 30 to 100 or 30 to 1000. In particular, the one-step accounting for clustering is consistently a good model, whilst in some situations the two-step and one-step ignoring clustering perform poorly.

Table 5. 9: Simulation results for the pooled effect size,$\widehat{\boldsymbol{\beta}}$, for three models of IPD meta-analysis, (two-step, one-step ignoring clustering and one-step including indicator variable), mean age is 4 with standard deviation 1.5 and the sample size with-in each study is between 30 to 100 observations, with the number of studies n=5 and 1000 simulations, the true values of the pooled effect size,$\boldsymbol{\beta}$, is shown in the table; the standard error for $\propto$ is 0.2 and 1.5.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (s. e. of $\propto$)	$\boldsymbol{\beta}$	Mean of $\widehat{\boldsymbol{\beta}}$	Bias	percentage bias*	MSE of $\boldsymbol{\beta}$	Percentage coverage	Mean s.e. of $\widehat{\boldsymbol{\beta}}$
1	5	Model 2	Two-step	-2.1(0.2)	0.30	0.30	0.00	-0.44	0.01	96.85	0.09
		Model 6	Ignoring clustering	-2.1(0.2)	0.30	0.30	0.00	1.41	0.01	96.29	0.09
		Model 8	One-step including indicator variable.	-2.1(0.2)	0.30	0.31	0.01	3.57	0.01	96.36	0.09
2	5	Model 2	Two-step	-2.1(0.2)	0.10	0.09	-0.01	-8.24	0.01	97.57	0.11
		Model 6	Ignoring clustering	-2.1(0.2)	0.10	0.09	-0.01	-5.95	0.01	96.58	0.11
		Model 8	One-step including indicator variable.	-2.1(0.2)	0.10	0.10	0.00	-3.67	0.01	96.58	0.11
3	5	Model 2	Two-step	-2.1(0.2)	0	0	0	NA	0.01	96.10	0.13
		Model 6	Ignoring clustering	-2.1(0.2)	0	0	0	NA	0.02	95.10	0.12
		Model 8	One-step including indicator variable.	-2.1(0.2)	0	0	0	NA	0.02	94.90	0.12
4	5	Model 2	Two-step	-2.1(1.5)	0.30	0.30	0.00	-0.95	0.01	95.40	0.10
		Model 6	Ignoring clustering	-2.1(1.5)	0.30	0.23	-0.07	-22.13	0.01	84.10	0.09
		Model 8	One-step including indicator variable.	-2.1(1.5)	0.30	0.31	0.01	3.53	0.01	94.80	0.10
5	5	Model 2	Two-step	-2.1(1.5)	0.10	0.10	0.00	0.35	0.01	97.70	0.11
		Model 6	Ignoring clustering	-2.1(1.5)	0.10	0.08	-0.02	-15.53	0.01	95.30	0.10
		Model 8	One-step including indicator variable.	-2.1(1.5)	0.10	0.10	0.01	5.45	0.01	96.10	0.11
6	5	Model 2	Two-step	-2.1(1.5)	0	0.00	0.00	NA	0.01	97.00	0.12
		Model 6	Ignoring clustering	-2.1(1.5)	0	0.00	0.00	NA	0.01	95.40	0.11
		Model 8	One-step including indicator variable.	-2.1(1.5)	0	0.00	0.00	NA	0.02	95.60	0.12

* Note that all the results are shown to two decimal places but the calculations such as bias, percentage bias, ... etc are based on full results (more than two decimal places).

Table 5. 10: Simulation results for the pooled effect size,$\widehat{\boldsymbol{\beta}}$, for three models of IPD meta-analysis, (two-step, one-step ignoring clustering and one-step including indicator variable), mean age is 4 with standard deviation 1.5 and the sample size with-in each study is between 30 to 1000 observations, with the number of studies n=5 and 1000 simulations, the true values of the pooled effect size,$\boldsymbol{\beta}$, is shown in the table; the standard error for $\propto$ is 0.2 and 1.5.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (s. e. of $\propto$)	β	Mean of $\widehat{\beta}$	Bias	percentage bias*	MSE of β	Percentage coverage	Mean s.e. of $\widehat{\beta}$
1	5	Model 2	Two-step	-2.1(0.2)	0.30	0.30	0.00	-0.08	0.00	96.57	0.03
		Model 6	Ignoring clustering	-2.1(0.2)	0.30	0.30	0.00	-0.39	0.00	96.36	0.03
		Model 8	One-step including indicator variable.	-2.1(0.2)	0.30	0.30	0.00	0.41	0.00	96.36	0.03
2	5	Model 2	Two-step	-2.1(0.2)	0.10	0.10	0.00	1.68	0.00	95.59	0.04
		Model 6	Ignoring clustering	-2.1(0.2)	0.10	0.10	0.00	1.72	0.00	95.68	0.04
		Model 8	One-step including indicator variable.	-2.1(0.2)	0.10	0.10	0.00	2.26	0.00	95.41	0.04
3	5	Model 2	Two-step	-2.1(0.2)	0.00	0.04	0.04	NA	0.04	94.90	0.04
		Model 6	Ignoring clustering	-2.1(0.2)	0.00	0.04	0.04	NA	0.04	95.30	0.04
		Model 8	One-step including indicator variable.	-2.1(0.2)	0.00	0.04	0.04	NA	0.04	94.80	0.04
4	5	Model 2	Two-step	-2.1(1.5)	0.30	0.30	0.00	0.22	0.00	95.40	0.04
		Model 6	Ignoring clustering	-2.1(1.5)	0.30	0.24	-0.06	-21.64	0.01	48.00	0.03
		Model 8	One-step including indicator variable.	-2.1(1.5)	0.30	0.30	0.00	0.74	0.00	95.30	0.04
5	5	Model 2	Two-step	-2.1(1.5)	0.10	0.10	0.00	-1.09	0.00	95.70	0.04
		Model 6	Ignoring clustering	-2.1(1.5)	0.10	0.08	-0.02	-19.03	0.00	88.90	0.04
		Model 8	One-step including indicator variable.	-2.1(1.5)	0.10	0.10	0.00	-0.60	0.00	95.50	0.04
6	5	Model 2	Two-step	-2.1(1.5)	0.00	0.00	0.00	NA	0.00	95.80	0.04
		Model 6	Ignoring clustering	-2.1(1.5)	0.00	0.00	0.00	NA	0.00	94.90	0.04
		Model 8	One-step including indicator variable.	-2.1(1.5)	0.00	0.00	0.00	NA	0.00	95.60	0.04

* Note that all the results are shown to two decimal places but the calculations such as bias, percentage bias, ... etc are based on full results (more than two decimal places).

5.5 The Impact of the Prevelance across Studies

In this simulation study, I did not examine the impact of the prevalence; that is the simulation design assumed the same prevalence of the PF in each study, and so relaxing this assumption may also effect on the models performance.

Altman et al.[136] states if there is variation in the prevalence across studies, then ignoring the clustering of patients across studies (and assuming the data as it came from one study) give an incorrect answer. They also stated that "*compared to standard meta-analysis, the treat-as-one-trial method gives greater weight to large trials and will tend to give narrower confidence interval"*.

To briefly look at how changes in prevalence in the PF across studies are creating differences in the models, I now present three datasets from my simulations where there are reasonable differences in the three model results. For dataset 3 (Table 5.11), the one-step ignoring clustering gets a smaller standard error and smaller effect size, as seen in the simulations. However the prevalence in each study is very similar, thus this shows that there can be differences in the model results even when the prevalence is the same in each study, the reason being that $\propto_i$ can still vary considerably which causes differences in model results. Similarly in dataset 2, the prevalence is similar but the standard error of $\hat{\beta}$ is much smaller in the one-step ignoring clustering.

In future work I will consider the additional impact of varying the prevalence on the IPD meta-analysis models. This may even increase the bias in the one-step model that ignores the clustering of patients across studies. In dataset 1, the prevalence are more different across studies, but the $\propto_i$ are also different, so it is difficult to tell if the prevalence is having an effect or not.

Table 5.11: Testing the impact of the prevalence on the pooled effect size of the PF.

	Meta-analysis models	$\hat{\beta}$ (s.e)	Study	Event/total	Prevalence	$\propto_i$
IPD	Two-step	0.56 (0.21)	1	52/254	0.20	-1.25
Dataset 1	One-step accounting for clustering	0.54(0.20)	2	57/354	0.16	-1.16
	One-step ignoring clustering	0.49(0.19)	3	19/113	0.17	-4.53
			4	56/306	0.18	-2.68
			5	12/44	0.27	-2.71
IPD	Two-step	0.02(0.15)	1	102/575	0.18	2.93
Dataset 2	One-step accounting for clustering	-0.02(0.15)	2	141/755	0.19	-1.74
	One-step ignoring clustering	-0.12(0.10)	3	122/598	0.20	-1.94
			4	100/477	0.21	-2.16
			5	206/999	0.21	-3.51
IPD	Two-step	0.13(0.17)	1	91/433	0.21	-.564
Dataset 3	One-step accounting for clustering	0.16(0.16)	2	26/122	0.21	-1.54
	One-step ignoring clustering	0.03(0.09)	3	89/426	0.21	2.94
			4	168/883	0.19	2.61
			5	184/925	0.20	-2.98

5.6 The Impact of the Correlation between $\propto$ and β Across Studies

Recall in chapter 4, I assessed whether the correlation between $\propto$ and β in IPD meta-analysis models will effect on the pooled effect size of the PF, by using the BFMA models. I showed mathematically and by empirical example that the correlation between $\propto$ and β has no impact on the pooled effect size of the PF, under a certain conditions: (i) the outcome and the PF are binary data; (ii) $cov(\hat{\propto}_i, \hat{\beta}_i) = -var(\hat{\propto}_i)$; (iii) and there is no heterogeneity across studies($\tau_\alpha^2 = 0$). But, this arise a question; if there is heterogeneity across studies($\tau_\alpha^2 \neq 0$), does the correlation between $\propto$ and β effect on the pooled effect size of the PF? To briefly look at this, I select one dataset from my simulation (dataset3, Table 5.11) to investigate. This dataset has a substantial variation for $\propto 's$, with $I_\propto^2 = 99.5\%$. I found the results from the BFMA model are exactly the same as two-step model (Table 5.11), that ignores the correlation. This indicates that there is still no impact for the correlation between $\propto$ and β, even when there is heterogeneity across studies. But this is just one example. In the future work, I will undertake a simulation study to examine the impact of the correlation between $\propto$ and β on the IPD meta-analysis model results more fully.

5.7 Limitation of My Work

This simulation study has limitations; firstly I only considered one variable (binary or continuous), i.e. a univariate model and I did not consider if there is a variation of the prevalence across studies, as discussed. Also, I specified a linearity assumption for the continuous variable. I focused on two-step approach (model 2) and one-step approach (model 6 and 8). Other models are available such as those described in Chapter 4, such as a bivariate model in the second stage of the two-stage method, and the models accounting for residual variation. Also, I specified no heterogeneity on the PF effect across studies, as in TBI data. But often heterogeneity will exist. Further simulations are required to see if my findings hold in situations with heterogeneity, and to evaluate the performance of the other models in Chapter 4, and to examine the impact of any variation of the prevalence across studies. Also, as in any simulation study, other parameter values could have been chosen. I based mine on the TBI dataset, but other values for the number of studies, $\alpha, \beta,$ prevalence, sample size etc are possible and may (or may not) reveal further findings. Also I have considered a binary outcome and fitted logistic regression models, the findings may (or may not) change if survival model were considered.

5.8 Discussion

In this chapter I have undertaken a simulation study to investigate the impact of ignoring the clustering of patients within studies and to compare one-step versus two-step models in IPD meta-analysis of PFs. I have undertaken the simulation study for small (30 to 100) and large (30 to 1000) sample size respectively, and for 5 or 10 studies in the meta-analysis of either a binary or a continuous PF with a binary outcome.

For many of the scenarios I assessed, the two-step approach, the one-step approach accounting for clustering, and even the one-step ignoring clustering give unbiased pooled

effect size estimates of the PF, and suitable coverage around 95%. However, the one-step ignoring clustering generates downward biased pooled effect size estimates of the PF when the standard deviation of $\propto$ is large (1.5), with a poor coverage less that 95% due to standard errors that are too small. When the number of studies in the meta-analysis was small (5), and the number of patients was small, the simulation results also show the two-step approach gives upward bias in some scenarios for the binary PF analysis, and gave coverage too large (>95%), due to large standard errors. However, the one-step model that accounts for clustering performed well in all scenarios considered.

Also, I selected three datasets from my simulation for the binary PF with reasonable difference in the three model results; to examine the impact of the prevalence and the correlation between parameters across studies. I found that when prevalence was similar, large differences can still exist due to the variability in$\propto$. Variability in the prevalence may increase model differences, but this is for further work.

Based on my results, I conclude that ignoring the clustering of patients within studies is not to be generally recommended, as it can cause too low a coverage and too small standard error (i.e. it is anti-conservative). My findings are echoed in non-meta-analysis settings too, and so this is my recommended model choice for IMPF projects. Peters et al.[137] undertook a comparison of methods for analysing cluster randomized trials, the authors concludes that ignoring the clustering for standard logistic regression model was highly anti-conservative. Bland[138] concludes that *'The effects of clustering can be large, inflating Type I errors'*, which is exactly the same as my findings here. Clark .[139] the author stated that *'standard errors are biased downwards when using single-level models even when there are as few as two observations per group on average'*. Hogan et al.[131] stated that *'Ignoring important sources of variation in any analysis can lead to incorrect confidence intervals and P values'*. Also,

Hetdker et al. [140] stated that '*traditional regression models can lead to biased estimates of uncertainty and different conclusions*' when clustering is ignored.

The results of one-step accounting for clustering and two-step approach are often very similar, only in a few situations with a binary PF did the two-step methods perform worst. This may be because the two-step model did not account for the correlation between parameters ($\propto_i$ and β_i) in the second step. Jones et al.[99] stated that *'Ignoring correlation can lead to different pooled estimates of the treatment difference and their standard errors'*. Mathew et al.[130] stated that '*We therefore re-iterate that our result is a theoretical result giving a complete description of the linear models under which the one-step and two-step IPD meta-analysis estimators coincide, assuming that the covariance matrices are known'* which suggest further that one-step and two-step meta-analysis approach may be coincide if we account for the correlation between parameters in the two-step, as then the whole covariance matrix of estimates is used (future work). Another key problem could be that the $\hat{\beta}_i$ estimates are assumed normally distributed in the second step, which may not be appropriate in situations where the sample size is small and the odds ratio is large. Other methods for dealing with pooling odds ratios may also overcome this issue as the one-step does[141 142].

Given this and that in more complex scenarios multiple PFs may be of interest in the same model, it seems sensible that the one-step method is the general choice for an IMPF. In chapter 6, I now consider application to the TBI data again and investigate non-linear trend in a PF effect.

CHAPTER 6

POSSIBLE IPD META-ANALYSIS MODELS OF CONTINUOUS PROGNOSTIC FACTORS BY USING FRACTIONAL POLYNOMIAL

6.1 Introduction

There are many issues for the analysis of continuous PFs, and a key one arises when the assumption of linearity is found to be untenable. Through my reviewing for the 20 IMPF articles (see chapter 2), only 6 articles considered a non-linear trend of the continuous PF; in particular, 4 of them used spline functions[75 77 76 78] and the other 2 used polynomial terms[72 80]. Though polynomial models are easy to fit, they may fit the data badly and give misleading inferences[143 144]. Fractional polynomials (FP) are a more flexible set of parametric models that can overcome the problems of polynomial models[145-147].

In this chapter, I develop meta-analysis models when the research interest is an estimate of the pooled effect of a continuous PF across studies, with potential non-linearity. In particular, the focus is on considering the possible two-step and one-step models for synthesising IPD studies that account for the clustering of patients across studies, whilst using first and second order FP terms within a logistic regression model. The TBI dataset again is used as an illustration to compare between one-step and two-step meta-analysis.

6.2 Fractional Polynomial Modelling on a Single Study

Regression models based on FP functions of a continuous covariate are described by Royston and Altman[143]. A FP function can be applied in various regression models such as simple and multivariable linear regression, logistic regression, and Cox proportional hazard models. In this chapter the focus is on using FP function within the context of IPD meta-analysis of a

PF, using logistic regression. Let us first consider how to apply FP in a single study using logistic regression.

The general m-degree FP function of a logistic regression model for a continuous variable, x (e.g. age), is given by

$$logit(P) = \beta_0 + \beta_1\ x^{(p_1)} + \beta_2\, x^{(p_2)} + \cdots \ldots\ldots + \beta_m\ x^{(p_m)} \qquad (6.1)$$

Where $p_1, p_2, \ldots \ldots p_m,$ denote fractional powers, m refers to the degree of the FP model, (the focus here is on $m \leq 2$ only), $x > 0$, and the round bracket notation denotes the Box-Tidwell transformation as follows:

$$x^{(p)} = \begin{cases} x^p & if\ p \neq 0 \\ \log(x) & if\ p = 0 \end{cases}$$

At first glance, it seems that FP model is similar to the family of conventional polynomial models; however it is more flexible as the powers, p, can be a non-integer values and the best fitted powers are to be determined. The powers, p, are chosen from a restricted set, S. Royston et al.[148] suggested $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, where x^0 denotes log (x); the set includes many possible transformations such as square root $(p = 0.5)$, reciprocal $(p = -1)$, logarithmic square transformations $(p = 0)$ and include no transformation (i.e. linear trend) when $p = 1$.

In this chapter, the focus is on a two degree FP function $(m \leq 2)$, as Royston et al.[148] stated that "*we have found that models with degree higher than 2 are rarely required in multivariable analysis. Fractional polynomials with* $m \leq 2$ *offer many potential improvements in fit compared with conventional polynomials of the same degree*". Also they stated that "*experience gained since 1994 has confirmed S as an excellent general choice*". In this chapter, I will use the same notation used by Royston and Sauerbrei [148]; in particular, the notation FP1 and FP2 refer to fractional polynomial with one power (p_1) and two powers

($p_{1,}p_2$) to represent the first and the second-degree FP models respectively. With the set S of powers just given, there are 8 models with FP1, 28 models with FP2 with distinct powers ($p_1 \neq p_2$) and 8 models for FP2 with equal powers ($p_1 = p_2$), note that in this case the FP2 models are in the form $logit(P_{ik}) = \propto + \beta_1 x^{p_1} + \beta_2 x^{p_2} \, logx$; so a total of 44 possible models.

6.2.1 Model Selection

Models with FP functions are generally fitted by using maximum likelihood estimation, including logistic regression[148]. The best-fitting model is the one whose power, p, gives the highest chi-square statistics or the smallest deviance[148] (minus twice the maximized log likelihood of a fractional polynomial model with power, p).

6.2.2 An Illustrated Example

I now give a simple example of FP modelling within a logistic regression model by applying to the first study of the 4 IPD studies within the TBI dataset that was introduced in chapter 1 (see Table 1.1), and analysed initially in chapter 4. Recall in chapter 4, I found that age was a PF for 6-month mortality but I assumed a linear trend for age. I also found that there was residual variation due to age, which suggests that the linear relationship may actually not be the most suitable.

The Best Fitted FP1 Logistic Regression Model

First I fit all of the possible 8 models with one power and decide the best fitted model by selecting the model that has the highest chi-square statistics or the lowest deviance.

Table 6.1 gives the results of the possible 8 models between age and logit probability of 6 month mortality. The best fitted model is the cubic model (model 8), as it has the lowest deviance (885.763) and the highest chi-square statistics (25.82) with one degree of freedom.

Model 8 indicates that there is a significant association between age^3 and the probability of six month mortality as the P-value < 0.001.

Table 6. 1: The results of possible logistic regression models for FP1, for the association between age (per 10 years) and the probability of six month mortality for the first study of TBI dataset

Model	Power (p)	Deviance	$\widehat{\alpha}$	$\widehat{\beta}$	S.e ($\widehat{\beta}$)	chi square statistic	p-value of $\widehat{\beta}$
1	-2	901.167	-0.755	-3.020	0.968	10.41	<0.001
2	-1	896.350	-0.237	-2.691	0.706	15.23	<0.001
3	-0.5	893.936	0.773	-3.357	0.813	17.65	<0.001
4	0	891.693	-2.299	1.000	0.227	19.89	<0.001
5	0.5	889.742	-3.209	1.139	0.245	21.84	<0.001
6	1	888.169	-2.200	0.311	0.064	23.41	<0.001
7	2	886.257	-1.684	0.041	0.008	25.32	<0.001
8	3	885.763	-1.506	0.007	0.001	25.82	<0.001

The model can be written as:

$$logit(P) = \widehat{\propto} + \hat{\beta}\, age^3 \qquad (6.1)$$

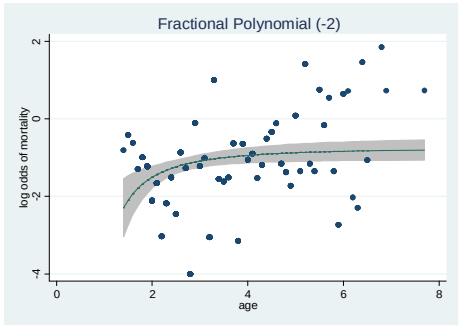
Where $\widehat{\propto}$ refers to the log-odds of the event (e.g. six month mortality rate) when $age^3 = 0$, age^3 is the cubic value of a patient age divided by 10, and $\hat{\beta}$ is the change in the log-odds of the event for one unit increase in age^3. There is, however, very little difference in the chi-square statistics value of the models with $p = 1,2, or\ 3$, suggesting little difference between a linear, quadratic and cubic. Note that the second best model is model 7, when $p = 2$ and the model is given as:

$$logit(P) = \widehat{\propto} + \hat{\beta}\, age^2 \qquad (6.2)$$

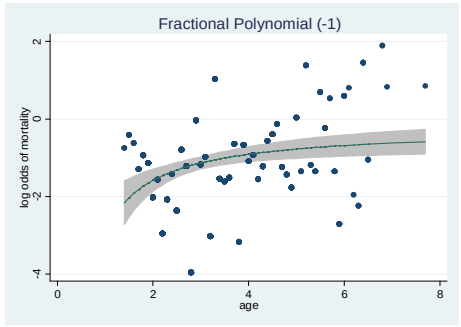
This model has the same interpretation as model 8, except that age^2 is the square value of the covariate (i.e. age); it has the second highest value of the chi-square statistics after model 8. Figure 6.1 shows the results of FP1 analysis of age, and the log-odds of mortality. In particular, the graph depicts the relationship between $age^{(p)}$ and log-odds of mortality based

Figure 6.1: All of the possible logistic regression models for FP1 logistic regression models to assess whether there is an association between age and in-odds of six month mortality for study one in the TBI dataset- the best model is $\boldsymbol{age^3}$ (Figure 6-1h)

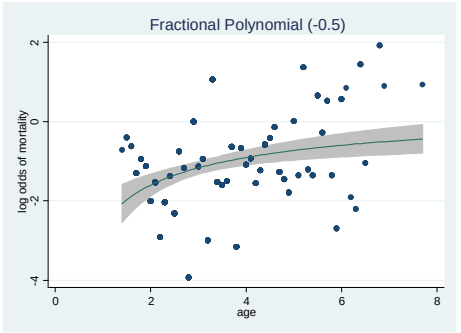
Figure(6-1a): $p = -2$



Figure(6-1b): $p = -1$



Figure(6-1c): $p = -0.5$



Figure(6-1d): $p = 0$

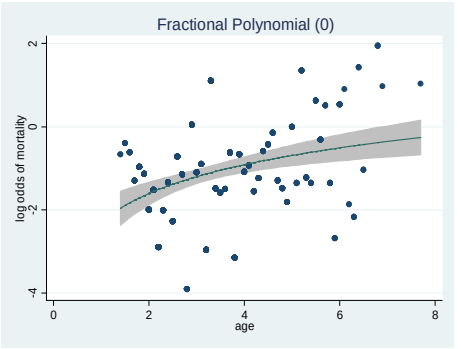
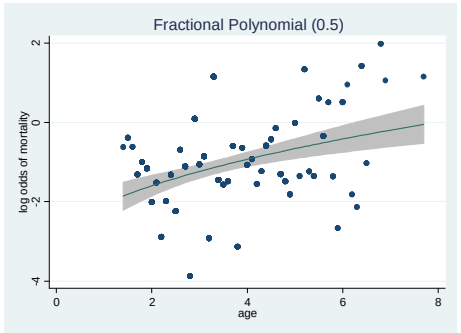
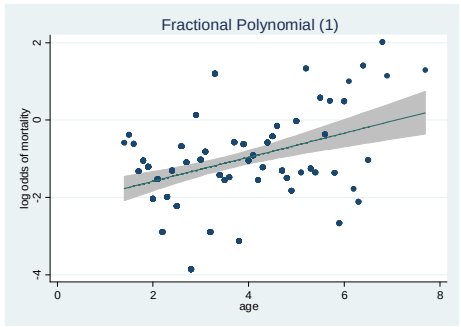


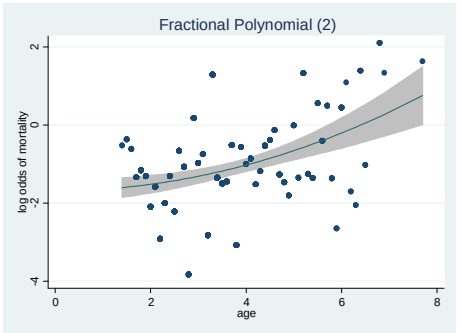
Figure (6-1e): $p = 0.5$



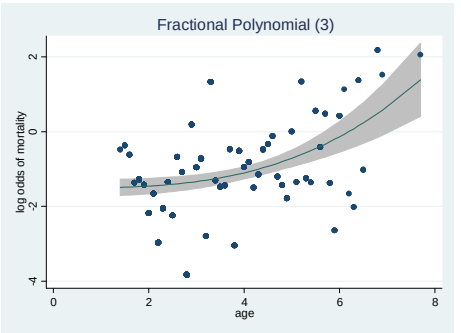
Figure(6-1f): $p = 1$



Figure(6-1g): $p = 2$



Figure(6-1h): $p = 3$



Note that STATA software is used to estimate these graphs; in particular "fracplot " module

on the logarithmic relationship fitted by logistic regression; shaded area denotes the 95% confidence interval around the line.

The Best Fitted FP2 Logistic Regression Model

FP2 is now applied to the first study in the TBI dataset, and thus 36 models fitted. Table 6.2 gives the results of the 36 models, the best model has powers ($p_1 = -2\,, p_2 = -1$), which has the lowest deviance and the highest chi-square statistics. The model is given as:

$$logit(p) = \hat{\propto} + \hat{\beta}_1 age^{-2} + \hat{\beta}_2 age^{-1} \qquad (6.3)$$

Here $\hat{\propto}$ refers to the log-odds of the event (e.g. six month mortality) when $age^{-2} = age^{-1} = 0$, age^{-2} is the inverse of the square value of the age, age^{-1} is the inverse value of age; $\hat{\beta}_1 \, and \, \hat{\beta}_2$ are the change in the log-odds of the event for one unit increase in $age^{-2} \, and \, age^{-1}$ respectively. Figure 6.2 shows the estimated relationship between age with power ($p_1 = -2\,, p_2 = -1$), and log-odds of mortality.

Figure 6.2: The best fitted logistic model by using FP2 , to investigate whether there is an association between age and log-odds of six month mortality, the best fitted model with powers $\boldsymbol{p_1 = -2 \, and \, p_2 = -1}$, for study 1.

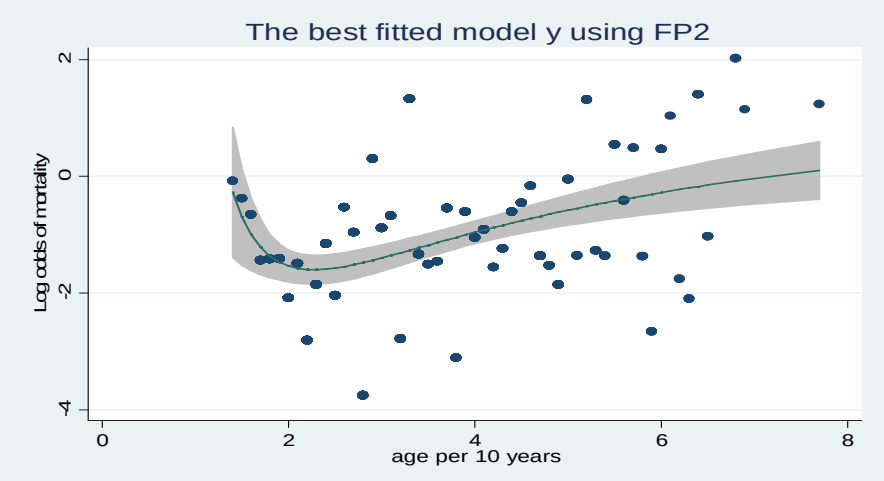


Table 6. 2: All of the possible logistic regression models by using FP2 to assess whether there is an association between age and ln-odds of six month mortality for one study of TBI dataset.

Power	Beta 1	Beta2	Beta1 s.e	Beta2 s.e.	chi square statistics	p-value 1	p-value 2	Power	beta1	beta2	Beta1 s.e	Beta2 s.e.	chi square	p-value 1	p-value 2
(-2,-1)	**17.68**	**-15.56**	**4.94**	**3.70**	**27.33**	**0.00**	**0.00**	(-1,2)	1.12	0.05	1.32	0.02	26.03	0.40	0.00
(-2,-0.5)	10.51	-12.06	3.29	2.88	27.27	0.00	0.00	(0,1)	-1.93	0.86	1.20	0.35	25.98	0.66	0.14
(-2,-2)	6.25	-21.27	2.30	5.05	27.26	0.01	0.00	(0.5,0.5)	-7.42	2.64	4.19	1.29	25.97	0.08	0.04
(-2,0)	6.90	2.56	2.50	0.62	27.14	0.01	0.00	(-0.5,2)	1.41	0.05	1.82	0.02	25.92	0.44	0.00
(-1,-1)	-2.00	-12.58	0.68	3.58	27.04	0.00	0.00	(-2,3)	0.38	0.01	1.24	0.00	25.91	0.76	0.00
(-2,0.5)	4.72	2.26	2.04	0.55	26.98	0.02	0.00	(0.5,1)	-4.12	1.39	2.62	0.69	25.87	0.12	0.04
(-1,-0.5)	22.08	-28.89	7.13	8.33	26.85	0.00	0.00	(0,2)	-0.46	0.06	0.63	0.02	25.84	0.47	0.02
(-2,1)	3.26	0.51	1.75	0.12	26.78	0.06	0.00	(-1,3)	0.10	0.01	1.09	0.00	25.83	0.93	0.00
(-1,0)	9.49	4.05	3.58	1.18	26.66	0.01	0.00	(0.5,3)	0.05	0.01	0.60	0.00	25.83	0.93	0.05
(-0.5,-0.5)	-14.45	-12.60	3.76	4.13	26.64	0.00	0.00	(1,3)	0.02	0.01	0.20	0.00	25.83	0.91	0.12
(-1,0.5)	5.29	2.95	2.42	0.87	26.48	0.03	0.03	(2,3)	0.01	0.01	0.05	0.01	25.83	0.92	0.48
(-0.5,0)	21.49	7.02	8.28	2.34	26.45	0.01	0.00	(-0.5,3)	0.02	0.01	1.41	0.00	25.82	0.99	0.01
(-2,2)	1.45	0.05	1.41	0.01	26.35	0.31	0.00	(0,3)	0.02	0.01	0.46	0.00	25.82	0.96	0.02
(-1,1)	3.20	0.59	1.86	0.18	26.31	0.09	0.00	(3,3)	0.01	0.00	0.02	0.01	25.82	0.66	0.98
(-0.5,0.5)	8.91	3.82	4.18	1.29	26.27	0.03	0.00	(1,1)	-0.94	0.54	0.81	0.35	25.80	0.25	0.12
(0,0)	-2.52	1.49	1.39	0.58	26.26	0.07	0.01	(0.5,2)	-0.63	0.06	0.92	0.03	25.79	0.49	0.05
(-0.5,1)	4.73	0.68	2.85	0.23	26.13	0.10	0.00	(1,2)	-0.24	0.07	0.37	0.05	25.76	0.51	0.13
(0,0.5)	-4.89	6.45	2.35	2.57	26.10	0.04	0.01	(2,2)	-0.02	0.03	0.09	0.05	25.76	0.83	0.51

I now consider how to apply a two-step or a one-step meta-analysis by using FP functions restricted to S subset.

6.3 Two–step IPD Meta-analysis

Recall that in the 'two-step' approach, the IPD are first analysed separately in each study using an appropriate statistical method for the type of data being analysed; for example, for binary outcome a logistic regression model might be fitted. This produces aggregate data for each study, such as the PF effect estimate and its standard error; these are then synthesized in the second step using a suitable model for meta-analysis of aggregate data, such as one weighting by the inverse of the variance whilst assuming fixed or random PF effects across studies (see chapter 1). Before fitting the two-step FP logistic regression model, testing of the linearity assumption can be considered for the relation between age and mortality rate by using closed test in each study.

Closed Test:

Closed test is used to choose the best FP function and it is also used to test the linearity assumption between the PF and outcome. In FP modelling the default function is linear $(p = 1)$. The procedure for closed test is given as follows[148-150], in relation to TBI data:

1. Calculate the deviance difference between the best FP2 model (i.e. the best fitted model with the lowest deviance) and null model (i.e. that which excludes the PF and has only the $\propto$ term) and test it at level $\propto$ on the chi-square distribution with 4 degree of freedom, If the p-value is non-significant, then stop and accept the null model, which assumes that the PF is not significant at the $\propto$ level, but if the test is significant, then go to the second step.
2. Calculate the deviance difference between the FP2 model and the linear model $(P = 1)$ at the $\propto$ level on the chi-square distribution with 3 degree of freedom. If the

p-value is non-significant, then stop and then accept the linear model (so in this case the linearity assumption is achieved); but if the p-value is significant this indicates that the linearity assumption is not achieved, so, go to the third step.

3. Calculate the deviance difference between the best FP2 model and the best FP1 and test it at the $\propto$ level on the chi-square distribution with 2 degree of freedom. If the p-value is non-significant then stop and FP1 is the best function, but if the test is significant, then the best function is FP2.

Note that step 1 tests whether the PF is significant or not; step 2 is used to test whether the linearity assumption between the PF and outcome is achieved or not; and step 3 compare between the simple and more complex non-linearity model.

Test of Linearity Assumption and the Best Selected Function

Table 6.3 gives the test of linearity assumption, and the best selected FP function for age per 10 years in TBI dataset for each study separately by using closed test. In the 4 studies, the linearity assumption is achieved for $\propto$ of 0.05, as the P-values for the 4 studies are not significant (Table 6.3). But, if I changed the level of $\propto$ to be, say, 0.10, then study 3 does not achieve the linearity assumption with P-value = 0.087, and the best fitted model is FP1 with $power = 3$ for this study. In this situation, I should decide whether I use linear or non-linear models for the 4 studies. Thus, the choice of $\propto$ is clearly important.

If I ignore the closed test, so the best fitted models are FP1 with powers 3, 1, 3, 2, for study 1, 2, 3, and 4 respectively. Then by considering the non-linear trend across the 4 studies, there are different powers across the 4 studies which make it difficult to apply a two-step model meta-analysis here, as it needs the 4 studies to have the same power (i.e. same model terms in each study). For example, if study one has βx^3 and study four has βx^2 , the β term cannot be synthesised. In the next section I give a suggestion on how to fit FP1 two-step IPD meta-

analysis model, by using logistic regression model, and how to decide upon which power to use consistently in each study.

Table 6. 3: Test of linearity and the best selected FP function for each study for TBI dataset in one-step IPD meta-analysis

Study	Model	Deviance	Power (p)	Steps of closed test	Comparison	Dev.diff	P-value	The best fitted model
Study (1)	FP2	884.246	-	1	FP2 vs. null	27.335	<0.001	
	Linear	888.169	1	2	FP2 vs. linear	3.923	0.270	
	FP1	885.763	3	3	FP2 vs. FP1	1.517	0.468	FP1=3
	Null	911.581	-			-	-	FP2=(-2,-1)*
Study (2)	FP2	1087.992	-	1	FP2 vs. null	28.844	<0.001	
	Linear	1088.920	1	2	FP2 vs. linear	0.928	0.819	
	FP1	1088.920	1	3	FP2 vs. FP1	0.928	0.629	FP1=1
	Null	1116.836	-			-	-	FP2=(3,3)*
Study (3)	FP2	565.171	-	1	FP2 vs. null	63.350	<0.001	
	Linear	571.751	1	2	FP2 vs. linear	6.579	0.087	
	FP1	565.492	3	3	FP2 vs. FP1	0.321	0.852	FP1=3
	Null	628.521	-			-	-	FP2=(0.5,2)*
Study (4)	FP2	425.401	-	1	FP2 vs. null	15.558	<0.001	
	Linear	426.049	1	2	FP2 vs. linear	0.648	0.885	
	FP1	425.795	2	3	FP2 vs. FP1	0.394	0.821	FP1=2
	Null	440.959	-			-	-	FP2=(-2,3)*

*Fitted FP1 logistic regression model is better than FP2 in each study of TBI dataset.

6.3.1 First Order Fractional Polynomial (FP1)

I now apply FP1 logistic regression models separately for each of the 4 IPD studies in the TBI dataset. Table 6.4 shows the chi-square statistics values for all possible 8 models. Figure 6.3 shows the fitted curve for each study for $p = 1, 2$ and 3. Figure 6.4 shows the comparison between FP functions for linear($p = 1$), quadratic ($p = 2$), and cubic ($p = 3$) function for each study separately for TBI dataset. Based on the closed test and $\propto = 0.05$ (i.e. significant level is 5%), all studies choose the linear trend as the best fitted relationship. In this case, one could apply the methods in chapter 4. However if $\propto = 0.1$, then study 3 rather chooses a cubic trend (Table 6.3) and so, there are different powers for the best fitted model across studies which make it difficult to apply a two-step meta-analysis, as mentioned. The difference of the 'best' model could be a reflection of between study heterogeneity in the true FP relationship between age and ln-odds of mortality, or it could be a consequence of chance variation. In

chapter 1, for the 4 TBI studies, recall that there was no heterogeneity when age was assumed to have a linear trend ($I^2 = 0\%$). Further, though the best fitted model for one study is different from the best fitted model in others, the visual difference between the fitted models with different powers in each study is not substantial (Figure 6.3 and 6.4). In particular, there is only a little difference between linear, quadratic and cubic functions for age especially when the age of patients is between 20 to 60 years in all studies; though there is a slight difference between the three functions when age is less than 20 years and bigger than 60 years respectively (i.e. at the extremes of the data with less values, especially for study 3).

Table 6. 4 The possible FP1 logistic regression models for each of the 4 TBI studies to assess the association between age and the probabilty of six month mortality

Study 1			Study 2			Study 3			Study 4			Total Chi-square for 4 studies
Model	Power	Chi-square	Model	Power	chi - square	Model	Power	Chi-square	Model	Power	Chi-square	
1	-2	10.41	1	-2	17.68	1	-2	27.07	1	-2	10.8	65.96
2	-1	15.23	2	-1	22.04	2	-1	36.68	2	-1	12.5	86.45
3	-0.5	17.65	3	-0.5	24.17	3	-0.5	42.12	3	-0.5	13.3	97.24
4	0	19.89	4	0	25.98	4	0	47.56	4	0	13.99	107.42
5	0.5	21.84	5	0.5	27.27	5	0.5	52.57	5	0.5	14.53	116.21
6	1	23.41	**6**	**1**	**27.92**	6	1	56.77	6	1	14.91	123.01
7	**2**	**25.32**	**7**	**2**	**27.2**	**7**	**2**	**61.97**	**7**	**2**	**15.16**	129.65
8	**3**	**25.82**	8	3	24.5	**8**	**3**	**63.03**	**8**	**3**	**14.89**	128.24

One could thus arguably choose either powers 1, 2, or 3. To help to identify the overall best model, I added the chi-square value across for all the 4 IPD studies for the same model (i.e. same power), to find the model with the overall highest chi-square statistics (Table 6.4). I found that the best FP1 model across all 4 IPD studies is model 7 with $p = 2$ (see Table 6.4),

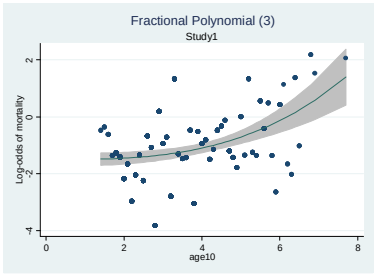
with total chi-square statistics equals 129.65. Thus, I decided to undertake the two-step IPD meta-analysis model for $p = 2$. This model is written as:

$$\begin{aligned} First\ step: \quad & logit(p_i) = \widehat{\propto}_i + \hat{\beta}_i age_{ik}^2 \\ Second\ step: \quad & \widehat{\beta}_i = \beta + e_i \ , e_i \sim N(0, Var(\hat{\beta}_i)) \\ & \hat{\beta} = \frac{\sum_{i=1}^{k} w_i \ \hat{\beta}_i}{\sum_{i=1}^{k} w_i} \end{aligned} \qquad (6.4)$$

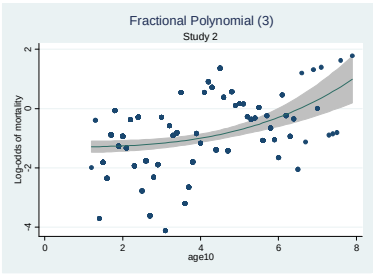
where $\widehat{\propto}_i$ represents the value of the log-odds when $age^2 = 0$, $\hat{\beta}_i$ refers to the change in log-odds of mortality for one unit change in age^2 in each IPD study; i refers to study. This allows a different $\propto$ in each study but assumes a fixedβ; random-effects on β are also possible if necessary. The pooled effect size estimate, $\widehat{\beta}$, gives the best estimate of the change in log-odds for one unit increase of age^2. Note that, for simplicity I did not consider the effect of the correlation between parameters α and β in the second step here (see further work, chapter 8). Figure 6.5 presents the forest plot for the association between age^2 and the probability of six month mortality. There is no heterogeneity across studies as $I^2 = 0\%$; thus the quadratic trend appears consistent in each study as did a linear trend (Chapter 4).

Figure 6.3: The fitted cubic, quadratic and linear regression model for 4 IPD studies of the TBI dataset

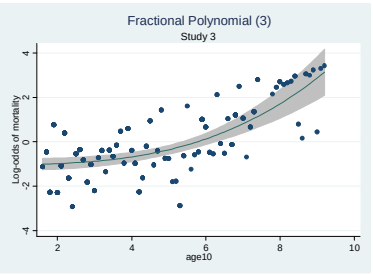
Figure(6- a):Study 1, $p = 3$



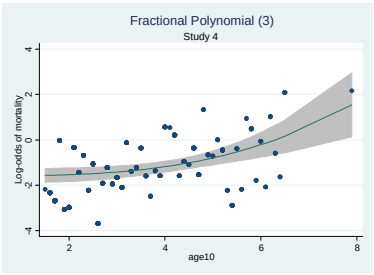
Figure(6- b):Study 2, $p = 3$



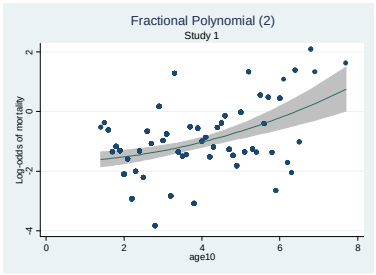
Figure(6- c):Study 3, $p = 3$



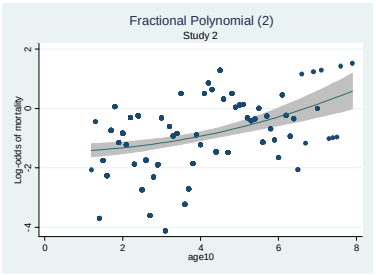
Figure(6- d):Study 4, $p = 3$



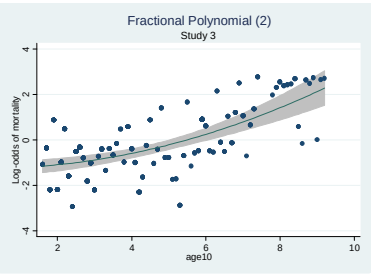
Figure(6- e):Study 1,$p = 2$



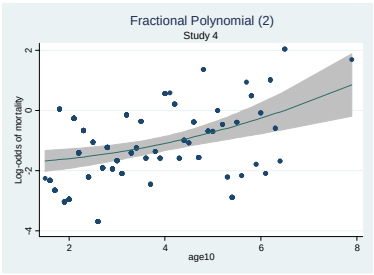
Figure(6- f):Study 2, $p = 2$



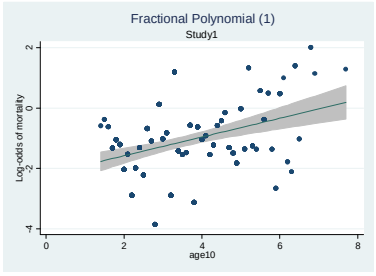
Figure(6- g):Study 3, $p = 2$



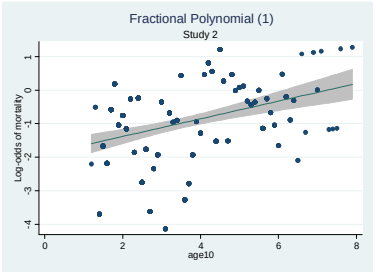
Figure(6- h):Study 4, $p = 2$



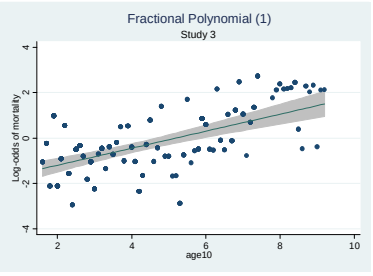
Figure(6- i):Study 1, $p = 1$



Figure(6- j):Study 2, $p = 1$



Figure(6- k):Study 3, $p = 1$



Figure(6- l):Study4, $p = 1$

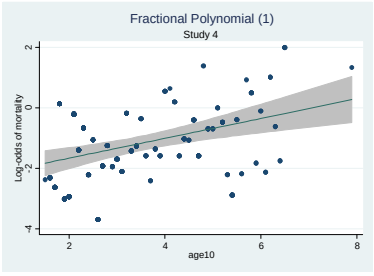


Figure 6.4: FP1 of linear, quadratic and cubic (p=1, 2, 3) models for age, for each study separately by applying on TBI dataset

Figure (6.4a): FP1 for linear, quadratic and cubic models for age for study 1, TBI dataset

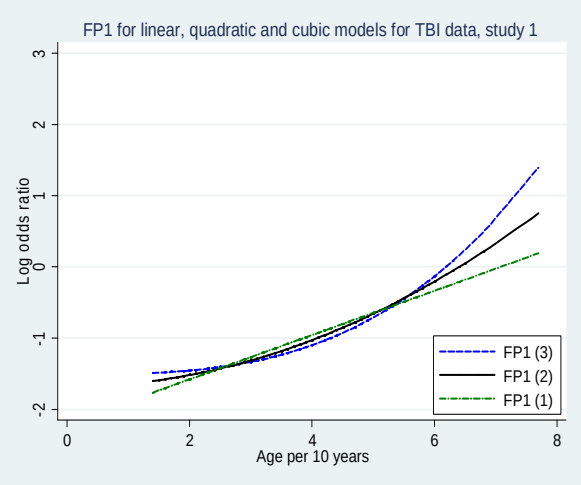


Figure (6.4b): FP1 for linear, quadratic and cubic models for age for study2, TBI dataset

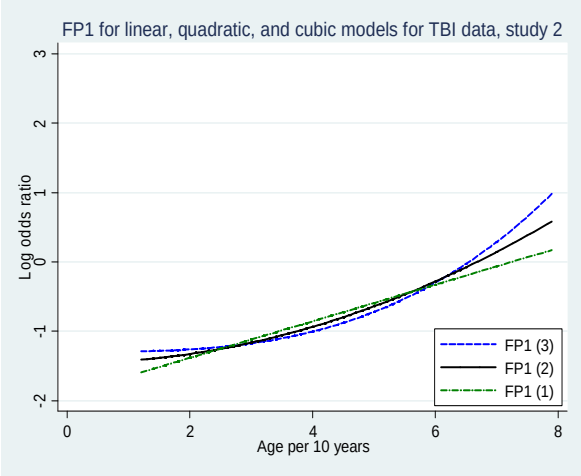


Figure (6.4 c): FP1 for linear, quadratic and cubic models for age for study 3, TBI dataset

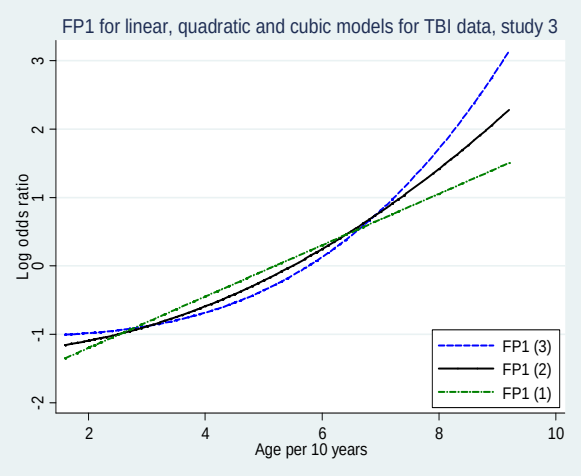


Figure (6.4d): FP1 for linear, quadratic and cubic models for age for study 4, TBI dataset

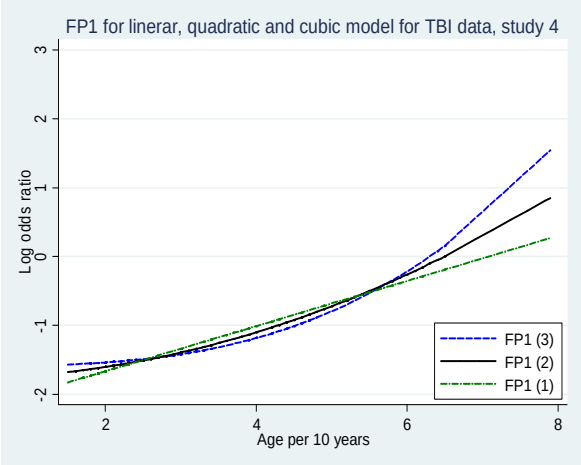
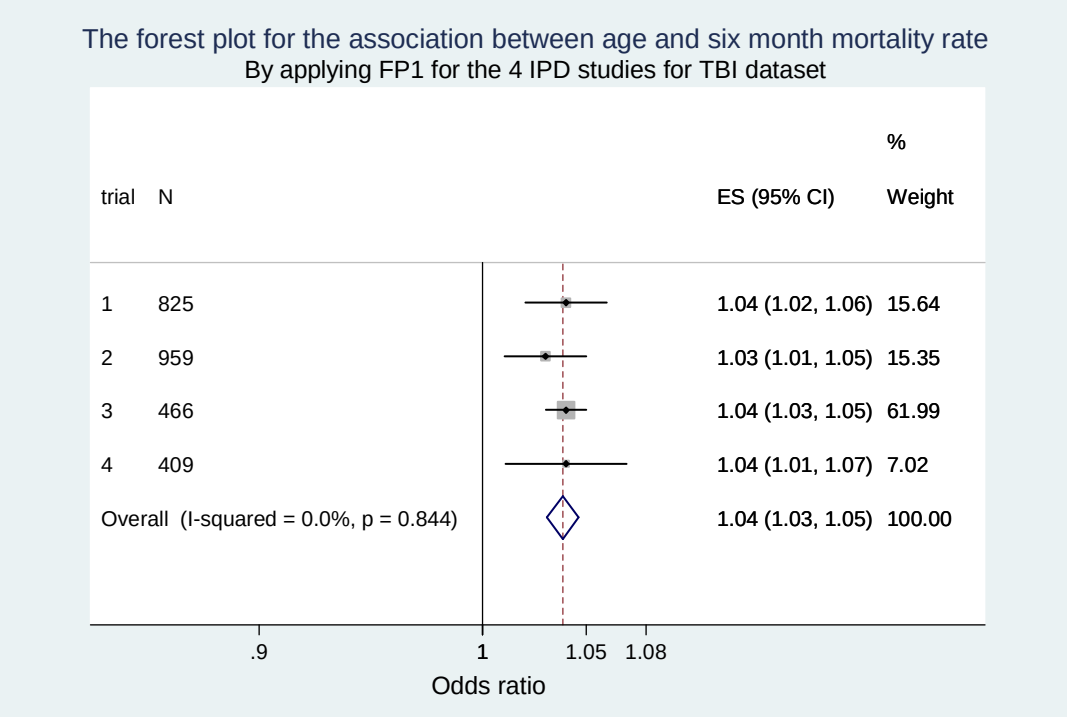


Figure 6.5: The forest plot for FP1 with power =2 logistic regression model of IPD meta-analysis by using fixed effect method- for the 4 IPD studies for TBI dataset



The pooled odds ratio for the 4 IPD studies is 1.04 with 95% confidence interval lies between 1.03 and 1.05; there is a significant relationship between age^2 and log-odds of mortality as one is not included in the confidence interval; the odds ratio (1.04) indicates that a one unit increase of age^2 multiplies odds of mortality by 4%.

Note that although FP2 function has the lowest deviance across the 4 studies, the closed test yielded that FP1 is better than FP2 as the reduction in deviance is not statistically significant at either the 5% or 10 % level. Thus, I will not consider FP2 two-step meta-analysis here. In situation where FP2 are required, it may be even more difficult to choose the same powers for all of the 4 studies in the meta-analysis, given there are 36 FP2 models possible in each study.

6.4 One-step Meta-analysis

Recall in the 'one-step' approach, the IPD from all studies are modeled simultaneously whilst accounting for the clustering of patients within studies. In this section I consider one-step

logistic regression meta-analysis models that uses the FP approach and account for the clustering of patients within studies. TBI is again used as an application. Models that ignore clustering are clearly not considered given the findings of Chapter 5. In the next sections, FP1 and FP2 models are fitted to select the best one-step IPD meta-analysis model.

6.4.1 First Order FP One-step Meta-analysis Model

In a one-step FP meta-analysis to adjust for clustering of patients within studies, I include study indicator as in chapter 4 (i.e. a separate intercept term); the choice of the first and second order FP model for the covariate of interest is then decided simultaneously across all studies (i.e. a separate FP model is not fitted per study). The FP1 one-step IPD meta-analysis logistic regression model which can be written as:

$$logit(P_{ik}) = \hat{\propto}_i + \hat{\beta}\, age_{ik}^{(p_1)} \qquad (6.5)$$

Where $\hat{\propto}_i$ is the log-odds of event in study i when $age_{ik}^{(p_1)} = 0$, β indicates the change in log odds for one-unit increase in $age_{ik}^{p_1}$, with β assumed fixed across studies, p_1 is the first order FP selected from the restricted S set (with 8 possible numbers).

One could also allow β to vary across studies, i.e. place a random-effects around it , either way, one is still forcing each study to have the same FP power terms, even if β varies. That is the choice of p is forced to be the same in every study. This is similar to the second step approach I used in section 6.3.1, where I forced the choice of p=2 in each study as my 'best' model.

6.4.2 Second Order FP for One-step IPD Meta-analysis

The FP2 one-step IPD meta-analysis logistic regression model is given as

$$logit\ (P_{ik}) = \propto_i + \beta_1 age_{ik}^{p_1} + \beta_2\ age_{ik}^{p_2} \quad (6.6)$$

Where α_i is the log odds of event in study i when $age_{ik}^{p_1} = age_{ik}^{p_2} = 0$, $\beta_1\ and\ \beta_2$ indicate the change in log-odds of the event for one-unit increase in age^{p_1} and age^{p_2} respectively; $age_{ik}^{p_1}$ and $age_{ik}^{p_2}$ are the fractional transformation of the covariate; again $p_1\ and\ p_2$ refer to the power of the continuous covariate and are selected from restricted S set for the best fitted model that has the highest chi-square statistics or the lowest deviance. Again the choice of p_1 and p_2 are forced to be the same in each study, even if the model is extended to allow heterogeneity in β_1 andβ_2.

6.4.3 Application to the TBI Data

Test of Linearity Assumption and the Best Selected Function

By applying the closed test on TBI dataset using the one-step approach, the linearity assumption is not achieved as the P-value for comparing FP2 versus linearity is significant ($p = 0.043$), which indicates that I go to step three to compare between FP2 and FP1 function. FP1 function is recommended over FP2 as the P-value for FP2 to FP1 is non-significant ($p = 0.793$), see Table 6.5.

Table 6. 5: Test of linearity and the best selected FP function for each study for TBI dataset in one-step IPD meta-analysis

Study	Model	Deviance	Power (p)	Step	Comparison	Dev.diff	P-value	The best fitted model
Study	FP2	2969.194	-	1	FP2 vs. null	128.702	<0.001	
(1)	Linear	2977.319	1	2	FP2 vs. linear	8.125	0.043	
	FP1	2969.657	2	3	FP2 vs. FP1	0.463	0.793	FP1=2
	Null	3097.897						FP2=(-2,2)*

*Fitted FP1 logistic regression model is better than FP2 in each study of TBI dataset.

Table 6.6: The results of two-step and one-step IPD meta-analysis models for TBI dataset by using FP1 and FP2

Model	Description of the Model	FP function	Chi^2	Deviance	p_1	p_2	Odds ratio β_1	Odds ratio β_2	95% confidence interval for Odds ratio β_1	95% confidence interval for Odds ratio β_2	Ln (odds ratio) β_1	Ln (odds ratio) β_2	95% confidence interval for Ln Odds ratio β_1	95% confidence interval for Ln Odds ratio β_2	P-value for $\widehat{\beta_1}$	P-value for $\widehat{\beta_2}$
Model1	Two-step	FP1			2	-	1.04	-	1.03 to 1.05	-	0.04	-	0.03 to 0.05	-	<0.00 1	-
Model 2	One-step Account clustering	FP1	585.99	2969.6	2	-	1.04	-	1.03 to 1.05	-	0.04	-	0.032 to 0.05	-	<0.001	-
Model 3		FP2	588.20	2969.2	-2	2	1.55	1.04	0 .44 to 5.42	1.03 to 1.05	0 .44	0.04	-0.82 to 1.7	0.03 to 0.05	0.495	< 0.001

N.B. all of results in this Table have been estimated by using STATA software, and the STATA code to fit the two-step and one-step models by using fractional polynomial are logit, fracpoly and metan, for more details see Appendix E.

I now apply the one-step FP models to TBI dataset and compare them to the second- step FP results. Recall that in two-step IPD meta-analysis models, I decided the best fitted model for FP1 was for age^2 (i.e. power = 2) based on the total highest chi-square statistics sum across studies; the odds ratio for a one-unit increase in age^2 was 1.04 with 95% confidence interval between 1.03 to 1.05. Using the closed test approach, for the one-step FP2 model, the best fitted model for FP1 is again age^2 (i.e. power = 2) (Table 6.5); and the estimates from one-step FP1 model are approximately the same as the two-step FP1 model (Table 6.6), and it has the same interpretation. In one-step FP2 model, the best fitted model for FP2 was $age^2\ and\ age^{-2}$, (i.e. powers 2, -2). However, the P-values were <0.001and <0.50 for $age^2\ and\ age^{-2}$ respectively. The insignificant P-value for age^{-2} suggest that $power = -2$ can be dropped from the model. Thus, the FP1 IPD meta-analysis model with age^2 again appears the most suitable model across all 44 choices.

6.5 Discussion

In this chapter I have proposed and illustrated statistical models for meta-analysis of binary outcome data in the presence of the one potential continuous PF with a potential non-linear trend, using two-step and one-step IPD meta-analysis models with first and second order FP function, whilst accounting for the clustering of patients within studies. Application was made to 4 IPD studies of TBI in relation to six month mortality. I found that it was more difficult to apply two-step methods, especially using FP2, as it is not straightforward to choose a single model with the same powers values in all studies. In contrast this problem does not arise when using one-step models as the powers in the model are selected using all studies simultaneously, forcing the same powers in each study, to automatically give an average trend.

The closed test for the same TBI dataset yielded different results when it is applied for the two-step or one-step method. For example, by assuming the level of $\propto$ is 0.05, the two-step method yielded that linear trend is achieved for the 4 IPD studies; when the level of $\propto$ was 0.10, it yielded that the linearity assumption is not achieved in study 3 of the TBI dataset. This makes it difficult to decide whether I consider linear or non-linear trend across studies in the meta-analysis. In contrast, in one-step approach, the closed test yielded nonlinear trend whether the level of $\propto$ is 0.05 or 0.10. Thus, this shows that one-step and two-step meta-analyses can differ in their conclusion, as shown in some simulations in Chapter 5.

Heterogeneity in the β terms can also be considered though I have not done that here. Thus I recommend one-step IPD meta-analysis approach for FP modelling as it is simpler and more coherent. Recall also from chapter 4 and 5 that the one-step also is unbiased, has suitable coverage and accounts for the correlation between parameters. In contrast, the two-step approach is occasionally biased with too high coverage and does not account for correlation unless the multivariate model is used.

Others have used FP functions in meta-analysis. In Rota et al.[151] the authors present a two-step meta-analysis using FPs. First, they examined how the risk of disease varies across different levels of a given exposure; they considered within studies variability, between studies heterogeneity, and nonlinear trend components; the second order fractional polynomial (FP2) generalised linear model is fitted for each study. Second, the pooled FP terms are estimated by using bivariate random-effects meta-analysis to consider also the heterogeneity across studies. They stated that "*The best fitting model, denoted by the optimal power transformations* (p_1^*, p_2^*) *is defined as the one minimizing Akaike's Information Criterion (AIC)*". So they decided the best fitted model in the second stage after they obtained the AIC for all 36 possible FP2 models that fitted for each study. It is not clear how they did

this, however, as I have shown in this chapter that the powers of the FP models might be different from study to study, making it difficult to choose the best FP model that have the same power in each study; Recall that I applied FP model for 4 IPD studies only. However, in Rota et al.[151] 14 case–control studies and one cohort study were included for the IPD studies; so I would imagine it was even more difficult to find the best fitted FP2 model in their example.

In Fedirko et al.[152] the authors aim to investigate whether alcohol consumption is related to colorectal cancer (CRC); Twenty-seven cohort and 34 case–control studies were used for investigate whether the dose-risk of alcohol increase the risk of colorectal cancer, the authors stated that "*Thirty-six second-order fractional polynomial random effects models and linear random effect models were tested.* And they also stated that "*The best-fitting model, defined as the one with the lowest Akaike's information criterion*". Again they did not mention how they select the best fitted FP2 (with the same powers for all IPD studies included in the analysis). However, they state that "*A random effects model was used to estimate pooled RRs in order to take into account the heterogeneity of the risk estimates and to provide more conservative estimates compared with the fixed effects model*". Greater clarity on how FPs is used in meta-analyses is needed in published articles.

FP modelling is also being used in meta-analysis when only aggregated data are available. In Bagnardi et al.[153] the authors synthesized the aggregated data (i.e. relative risk) for the second order fractional polynomials (FP2). In particular, those papers that investigate the association between alcohol intake and relative risk of all-cause mortality reported from 29 cohort studies. The FP powers are different from study to study; and the authors did not say how they overcome this problem and how they select the best fitted fractional polynomial with the same powers from the 36 models across 29 articles, they only stated that "*The best-fitting*

model $(p1 = p2 = 0.5)$ *offers a gain in deviance of 138.36 with respect to the reference model across the 29 cohort studies"*.

In conclusion, further research into fitting FP models within IPD meta-analysis is needed to consider one-step and two-step FP models, and how to select the best fitted model. But, based on my findings of this chapter and chapter 5, I would suggest a one-step meta-analysis framework for FP modelling should be preferred, which also accounts for the clustering of patients within studies. However a limitation is that it forces the same shape in each study. If this seems unrealistic, then it may not be sensible to do meta-analysis or then a two-step approach of FP is about to be suggested by Sauerbrei et al. that allows for different powers across studies (personal communications) followed by a synthesis.

CHAPTER 7

SMALL-STUDY EFFECTS IN IPD META-ANALYSIS OF PROGNOSTIC FACTOR STUDIES

7.1 Small-study Effects and publication related biases

Meta-analysis of PF studies that use aggregated data are known to be highly subject to publication bias and selective reporting biases[7 22]. Though meta-analyses of individual patient data (IPD) are often considered as the "gold standard" method in medical research for summarizing quantitative evidence, there has been little research into whether they too are affected by bias[7 22]. This is particularly important for IMPF as primary PF studies are prone to selective and biased reporting, and IPD may only available from published PF studies[7] (see chapter 2). In this chapter I perform an empirical assessment of whether there is potential bias within existing IMPF projects; in particular whether smaller studies have the tendency to produce more favourable PF results (small-study effects).

7.1.1 Publication Bias

Publication bias occurs when studies with statistically significant results are more likely to be published compared to studies with non-significant results; if IPD are only obtained from published studies, then publication bias remains a concern from IPD meta-analysis just as for an aggregated data meta-analysis[104 154]. Publication bias may arise from the design of the studies. For example, studies with inadequate sample sizes are unlikely to demonstrate statistical significance for clinically important effects, and may lead to publication bias if the results from small studies are unlikely to be published unless significant. This is referred to as 'small-study effects', the tendency for smaller studies to be missing or only published when they show significant results.

In addition, some editors and reviewers tend to dislike negative studies[104 155 156]. The *British Medical Journal* stated that "*negative results have never made riveting reading*.", as they would like articles to be published that affect clinical practice or improve prognosis; as a result of that the researchers may decide not to submit their negative findings, or hide negative findings and submit only the positive findings in their research[154].

7.1.2 Availability Bias

In an IPD meta-analysis, another related problem is availability bias, when not all known studies provide their IPD; this potentially leads to biased set of studies being included in the IPD meta-analysis. For example, in Trivella et al.[72] the authors asked for IPD from 38 studies but only 18 studies gave their IPD; the remaining 20 IPD studies that did not provide their IPD may cause availability bias if their results systematically differ. There are many potential causes for unavailable IPD[53 55 60]; such as there being no access to the data; authors being worried to give their data because they are not convinced about their statistical analysis or results; others may not like to share their data if the result is statistically non-significant; others may not provide their IPD if they do not have the complete data for particular confounding factors that were asked for or a lack of availability of information on desired outcomes; and others may have destroyed the data or refuse to provide it because they spent a lot of money on their research[53 55 157]. Availability bias can be investigated by combining the aggregated data from the IPD studies with that extractable from other non-IPD studies, to understand the possible impact of non-IPD studies on the results and conclusion[55 60 154].

7.1.3 Aims of the Chapter

Publication bias and availability bias are particular concerns for an IMPF articles; recall that in chapter 2 I reviewed 20 IMPF articles, and only 2 articles[72 93] included unpublished studies (which raises the threat of publication bias), and only 2 articles[80 88] obtained IPD from all

studies that they asked for (which raises the threat of availability bias). Further, only 2 articles[83] [92] considered the issue of publication bias by using funnel plot or Egger's test, and only one article[81] considered the problem of availability bias by examining whether including non-IPD studies will potentially affect on the IPD meta-analysis results. Thus the question arises: are IMPF articles subject to publication and other biases that may cause small-study effects?

This chapter aims to assess small-study effects in the 20 IMPF articles considered in detail in chapter 2, by using visual and statistical tools such as funnel plot[158], trim and fill method[159], cumulative meta-analysis[45] and tests for small-study effects (e.g. Egger regression test)[159]. These methods are now introduced.

7.2 Methods to Assess Small-study Effects

Checking for evidence of small-study effects should be undertaken routinely in a meta-analysis whenever possible[20]. In this section, four existing methods are described for detecting small-study effects: funnel plot, cumulative meta-analysis, trim and fill, and regression methods. They are applied again to the IPD from Look et al. [14] for illustration. Recall the Look data was introduced in section 1.4.2 and first analysed in chapter 2. A general rule of thumb is for at least 10 studies to be available in order to reliably use these methods[160].

7.2.1 Funnel Plot and Contour-enhanced Funnel Plot

A funnel plot is the most common method for detecting small-study effects and potentially publication bias[161]. A funnel plot displays the effect estimates (x-axis) against the standard error (y-axis)[158], and it places the most powerful studies in the top of the graph[20 160]. In the absence of publication bias, the plot is expected to be symmetric (PF effect estimates are scattered around the true effect size of the PF in both directions with increasing variability as the standard error increases). However, if small-study effects exist this changes the shape of the

plot and asymmetry will occur as some small studies will be missing in one direction[154]; in particular when there is potentially publication bias only those small-studies yielding a large effect size of the PF will be included. This indicates that there is an association between the sample size and the available PF effect estimates.

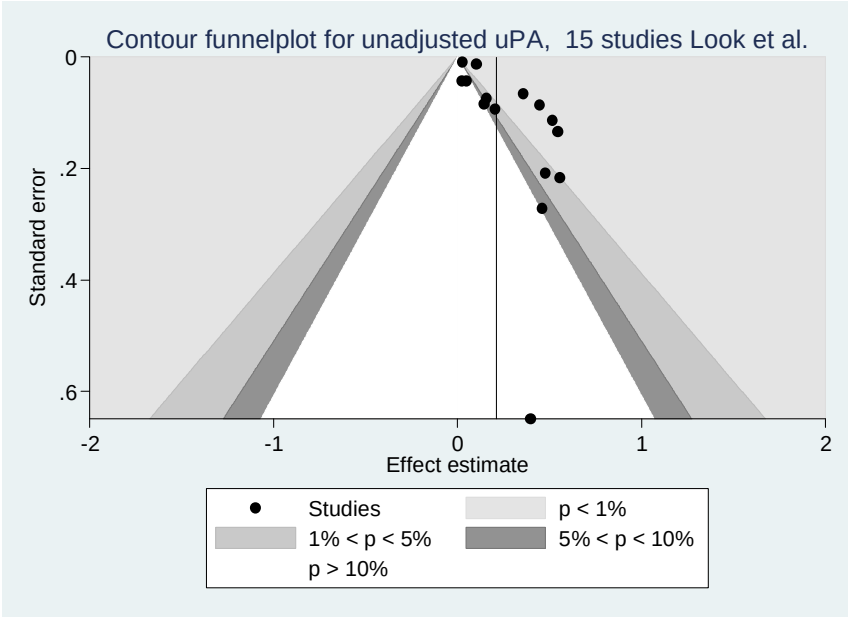
Note that there are many factors that may cause asymmetry in a funnel plot, not just publication bias, but also poor methodological quality (e.g. poor methodological design and inadequate analysis in small studies), and between-study heterogeneity in the true PF effect. Recent recommendations[160] note that to test the asymmetry by using funnel plot the number of the studies should be at least 10, with a substantial difference for the standard errors of the effect size estimates of the PF across studies (i.e. if the standard errors of the effect size estimates of the PF are similar, the funnel plot should not be used) as 'small' studies cannot be distinguished from other studies. Funnel plot can show whether there is asymmetry across studies or not, but it cannot show the reasons for this asymmetry[158 162].

A contour enhanced funnel plot is a funnel plot that include contour lines corresponding to perceived 'milestones' of statistical significance ($p = 0.01, 0.05, 0.1$ etc)[163-164]. This allows the statistical significance of the PF effect size of the PF in the seemingly missing studies to be considered; also, it may help to identify whether the asymmetry is due to publication bias or from other factors. For example, if the 'missing' studies are seemingly in the non-significant area ($P > 0.05$) of the contour-enhanced funnel plot, this possibly indicates that the asymmetry is due to publication bias. In contrast, if the missing studies are in the significant area of the plot (i.e. $P < 0.05$), this may indicate that the asymmetry comes from other factors (e.g. heterogeneity, poor study design, etc).

Figure 7.1 shows the contour funnel plot for the Look et al.[14] database, relating to the log hazard ratio and its standard error for uPA as a PF for relapse-free survival for breast cancer.

Note that uPA was measured as a continuous variable. It can be seen from the graph that most of the studies provide log hazard ratio estimate greater than zero, but the larger studies are those with a log hazard closest to zero. There is visual asymmetry, with small studies with results close to or less than zero potentially missing, broadly in the non-significance area (in particular in the white area where $p > 0.1$). This makes small-study effects a concern here, though heterogeneity for uPA ($I^2 = 83.4\,\%$), or other factors may also be the cause of asymmetry here (e.g. it may also come from the overestimation of PF effect in small IPD studies of inadequate methodological quality).

Figure 7.1: Contour funnel plot for log hazard ratio of uPA and its standard error for relapse-free survival (RFS) as outcome with 15 IPD studies from Look et al.[14]



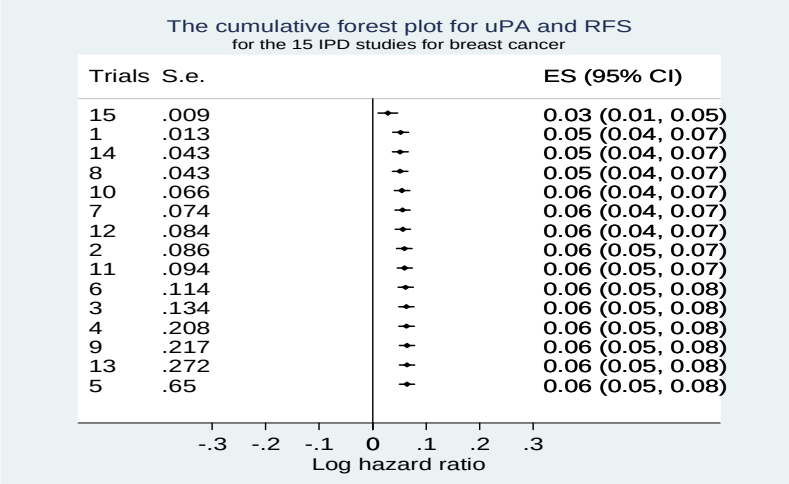
Black line is the meta-analysis pooled result.

7.2.2 Cumulative Meta-analysis

A cumulative meta-analysis is a method to assess the impact of additional studies to the meta-analysis. Essentially, it starts with the estimate of the effect size of the first study, then does a meta-analysis to estimate the pooled effect size for the first two studies, then repeats this for the first three studies, and so on[113]. The same thing will appear in the cumulative funnel plot,

the first line shows the effect size and the confidence interval for the first study, the second line shows the meta-analysis effect size estimate and the confidence interval for the first two studies together, and so on[113]. The method was originally developed to look at the impact of new studies over time, but I will use it here to examine the impact of smaller studies on the IPD meta-analysis. The studies can be sorted by the sample size (from high to low) or by the standard error (from low to high) [113]. I will sort the IPD studies according to their standard errors (from low to high). First, I start with the estimate of the PF effect size for the lowest standard error followed by a meta-analysis estimate of the two lowest standard errors and so on. If the cumulative meta-analysis effect size has stabilized with the inclusion of the lowest standard error and does not substantially drift with the addition of the smaller studies, this indicates that the smaller studies have only a little impact on the pooled effect results and thus small-study effects are not a big concern. In contrast, if the cumulative effect size of the PF drifts after adding the largest standard errors, this indicates that the smaller studies do effect on the meta-analysis effect size for the PF and suggests small-studies effects and potential publication bias.

Figure 7.2 Cumulative meta-analysis plot based on log hazard ratio and its standard error for uPA and relapse-free survival by using the random- effects method.



S.e: refers to the standard error for each IPD study and it is sorted from low to high, and Trials: refers to the study number according to the order for the standard error (from low to high)

Figure 7.2 shows the cumulative forest plot for the log hazard ratio of uPA for the 15 IPD studies from Look et al.[14] In particular, the graph shows that the IPD studies were sorted according to standard error (from low to high); the effect size of the PF drifted slightly to the right hand side of the graph. This shift, however, is not substantial as the pooled effect size is changed slightly from 0.05 to 0.06 when the standard error was 0.013 to 0.65 respectively. This indicates that there is a slight concern for the small-study effects in uPA, but it is potentially not substantial.

Note that as contour funnel and cumulative meta-analysis plots are visual tools, they may be considered as subjective methods, so statistical tests for funnel plot asymmetry can be helpful to quantify asymmetry as now discussed.

7.2.3 Egger's test for asymmetry

Recently, regression based adjustment methods have been used to assess the effect of publication bias on the pooled effect[165]. Such regression methods are used to detect small-study effects and adjust the pooled effect size of the original meta-analysis estimate of the PF according to that potential bias; these methods are now explained.

Eggers Methods

Egger et al.[106] proposed a regression test for funnel asymmetry that is widely used. There are two methods of Egger test. The first method is original Egger test and the second one is called Egger D-var method. These two methods are explained as follows:

(1). Original Egger Method

This method is equivalent to a variance-weighted meta-regression model [165]. This is given as

$$\widehat{\beta}_i = \propto + \beta \times se_i + \varepsilon_i \ \ weighted\ by\ \frac{1}{se_i^2}\ with\ \varepsilon_i \sim N(0, se_i^2 \times \emptyset) \qquad (7.1)$$

where $\widehat{\beta}_i$ is the effect estimate of the PF for each study i (i.e. log odds ratio or log hazard ratio from study i), and se_i is the standard error for $\widehat{\beta}_i$; α and β represent the adjusted pooled effect of the PF (intercept) and the slope associated with funnel plot asymmetry respectively. The regression is weighted by the inverse variance for each study i. $\emptyset$ refers to an unknown multiplicative dispersion parameter that allows for possible heterogeneity. Two key results from this model are $\widehat{\propto}$ which is the pooled effect size having adjusted for asymmetry[165], and the P-value associated with $\hat{\beta}$, which gives Egger's test for asymmetry[106].

(2). Egger D-var method

This method is similar to the original Egger method except that it replaces the standard error for the effect size of the PF for each study i (se_i), with the corresponding variance. This indicates that the relationship between the effect size of the PF and its variance is linear. Note that Egger D-var model is identified by Moreno et al.[165] as the method with potentially most appealing statistical properties for small-study effects adjustment[165].

$$\widehat{\beta}_i = \propto + \beta \times se_i^2 + \varepsilon_i \ weighted\ by\ \frac{1}{se_i^2}\ with\ \varepsilon_i \sim N(0, se_i^2 \times \emptyset) \qquad (7.2)$$

The parameters here have the same interpretation in equation (7.1), except that se_i^2 is the variance for each study i.

Moreno et al.[165] have undertaken a simulation study for the Egger's method in relation to obtain adjusted estimates (i.e. the adjusted estimates of the effect size of the PF). The authors found that if there is small number of studies in the meta-analysis and (or) heterogeneity across studies, original Egger's methods yielded too high coverage, low residual bias and high MSE values (due to large variance), while Egger's D-var method yielded lower, more suitable coverage and lower MSE. Based on this the authors conclude that "*we recommend Egger's D-var method which perform very similarly throughout the simulations; at least in*

terms of coverage, MSE and variance". So, I will consider the Egger D-var method for the adjusted pooled effect size of the PF in this chapter.

Applying to Look et al[14].

I applied the Egger's regression methods to the Look et al.[14] data. Egger's test for asymmetry gave P-value =0.002 so there is evidence of small-study effects. The original meta-analysis for the hazard ratio of the PF was 1.22 with 95% confidence interval (1.15 to 1.31). As there is heterogeneity (I^2 =87.4%), I applied Egger's D-var method and the adjusted pooled effect size of uPA is 1.06 with 95% confidence interval (1.02 to 1.11). This still indicates that uPA is a PF (same as the original data results); although the pooled result is clearly much close to 1. Thus even though small-study effects are a slight concern given the significant asymmetry, uPA appears to still be prognostic even after adjusting for this asymmetry.

7.2.4 Trim and Fill Method

The trim and fill method is another method to assess publication bias and small-study effects. Peters et al.[166] state that *"the trim and fill method is an iterative non-parametric method based on the asymmetry of a funnel plot to estimate an adjusted pooled effect size; and it is recommended to be used as a form of sensitivity analysis of the pooled effect size"*. The trim and fill method aims both to identify and correct for funnel plot asymmetry arising from small-study effects[20]. The main idea of the trim and fill method is; (i) to trim the small studies (usually on the right hand side for PF studies) that potentially have an extreme effect size of the PF which caused asymmetry[159] (ii) after omitting the small studies (which might lead to overestimated of the pooled effect size), the pooled effect size of the PF is re-estimated again[106]; (iii) estimate the missing studies on the left-hand side that reflects the omitted studies on the right-hand side of the funnel plot based on symmetry about the pooled effect from (ii); (iv) the adjusted pooled effect size estimate of the PF is calculated by performing a meta-

analysis including these 'filled' studies[159 166]. This process is repeated until estimates of the number of missing studies and pooled effect size of the PF are stable[166].

Estimators of the number of missing studies

When there is no publication bias, the number of the individual studies equals n (i.e. the number of studies in the original meta-analysis). For each $i = 1$ to n study, study i produces an effect size β_i which estimates β (the average PF effect), and an estimated within-study variance[113], σ_i^2 . If the problem of publication bias exists, the true number of the individual studies is $n+ k_0$; k_0 refers to the missing studies that should also be included in the estimation of β[113]. Three estimators for k_0 are described by Duvel and Tweedie[159 166]: L_0, R_0, and Q_0, but they conclude that L_0 and R_0 perform better than Q_0 and so are the preferred estimators.

In this chapter the focus is on L_0 estimator; for more details about the other estimators, see Duvel and Tweedie.[104 159 167] the L_0 estimator is given as:

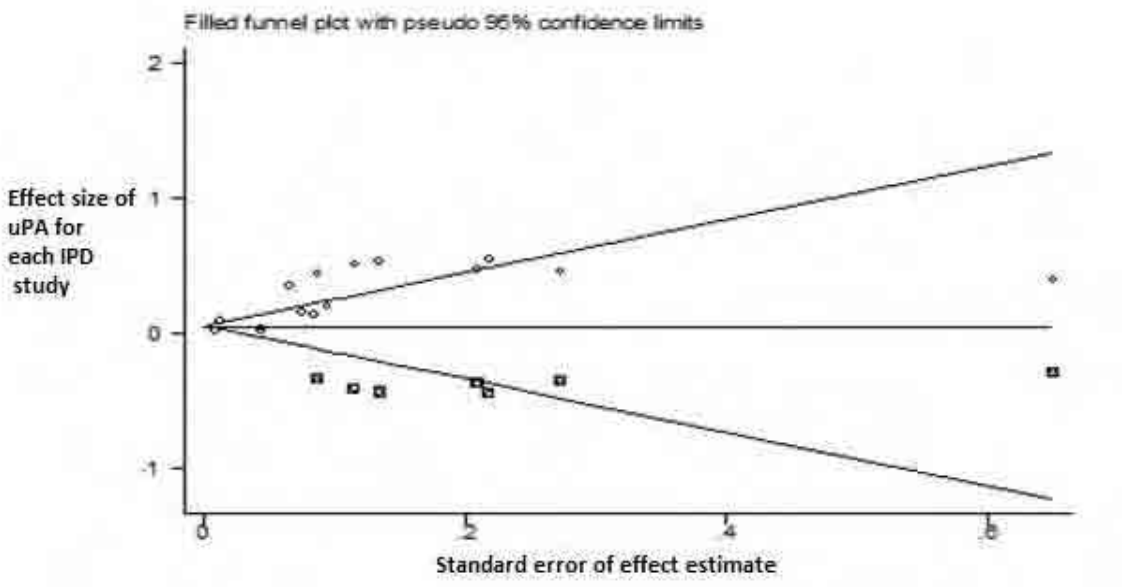
$$L_0 = \frac{4\,S_{rank} - n(n-1)}{2n-1} \tag{7.5}$$

where n is the original number of the studies in meta-analysis, S_{rank} is the Wilcoxon statistic (see Rothstein et al.[104] for how to calculate S_{rank}). Note that at each iteration L_0 is rounded up to the nearest integer and provides estimates of k_0. After estimating the number of missing studies, the meta-analysis is repeated on the filled dataset to get "trim and fill" estimate (i.e. the adjusted pooled effect size). Each filled study is assigned the same standard error as the trimmed study it reflects in order to maintain symmetry within the filled dataset.

Note that Peters et al.[166] conclude that if there is evidence of publication bias in the effect size with large between study heterogeneity, the fixed-random trim and fill method performing better than fixed-fixed and random-random trim and fill method. They state that "*the fixed–fixed effects trim and fill models give estimates with very narrow 95 percent CIs.*

Given that between-study heterogeneity is present, these 95 percent CIs are likely to be too precise as they do not account for this between-study heterogeneity. The fixed–random effects trim and fill models give wider CIs than the fixed–fixed effects model, since a random effects model is used to synthesize the filled data set, but the random–random effects model gives the most conservative CIs". Thus, in this chapter, the fixed-random trim and fill method is performed to estimate the number of missing studies.

Figure 7.3: Trim and fill plot based on the log hazard ratio and its standard error for the association between RFS and increase of uPA in Look et al.[14] after filled by the effect size of the missing studies by using fixed- random method



I applied the fixed-random trim and fill method to the Look et al.[14] dataset. Figure 7.3 shows the trim and fill plot for the log hazard ratio and its standard error for the uPA. As can be seen from the graph, there are seven missing studies imputed indicated by a square around the data symbol. Again, the original random-effects meta-analysis is 1.22 with 95% confidence interval (1.15 to 1.31). The adjusted pooled effect size from fixed-random trim and fill method after imputing the 7 missing studies is 1.10 with 95% confidence interval (1.03 to 1.17). This leads to the same conclusion as the adjusted result when using Egger D-var; that is although there is some potential evidence of small-study effects, uPA is still a PF.

7.2.5 Summary for small-study effects in Look et al.[14]

In this section, I summarize the findings results for uPA for the Look et al.[14]. The original meta-analysis of the PF was 1.22 with 95% confidence interval (1.15 to 1.31). The contour funnel plot shows asymmetry, and potentially missing studies appear to lie in the white area of the graph (insignificant area). This indicates that small-study effects are a potential threat, and publication bias may be operating. There is also a slight shift to the right hand side of the cumulative forest plot as small-studies are introduced. The Egger's P-value is 0.002, and thus significant. The trim and fill method yielded 7 missing studies. So it is of interest to use Egger's D-var method to obtain an adjusted estimate of uPA. It yielded a change in their pooled estimate value compared to the original meta-analysis method (see section 7.3.3). However, uPA is still a PF, and this does not change the clinical conclusion of the original results that uPA is prognostic.

7.3 In-depth Assessment of Small-study Effect in the 20 IMPF Articles

As mentioned, small-study effects were rarely assessed in the 20 IMPF articles examined in Chapter 2, so I now do this using the methods introduced in section 7.3.

7.3.1 Obtaining Suitable Data from 20 IMPF Articles

To examine small-study effects, I needed to obtain the PF effect estimate and its standard error for each IPD study in each IMPF article. Unfortunately, I only found one article that reported the PF effect size and its standard error (or the confidence interval that I then used to calculate the standard error) for each IPD study, this was Trivella et al.[72]

In the other IMPF articles, the authors gave the pooled effect size of the PF and its standard error, but without study specific estimates; for example the forest plot visually showed the effect size of the PF and its confidence interval for each IPD study but without their exact

values given. This did not enable me to extract the data required, unless potentially inaccurate extrapolation was made from the forest plot.

To help overcome this, I e-mailed the authors of the other 19 IMPF articles to ask them for the aggregated data for their PF studies; in particular, the effect size of the PF and its standard error (or its confidence interval) for each IPD study. I also asked for the IPD itself (if it was available). The authors of the IMPACT study kindly agreed to give me their IPD data. Of the 13 other studies, only five authors replied, and only the 2 MeRGE articles[92 93] provided their aggregated data. Thus, in the next section I examine whether small-study effects appear to exist and their impact in the Trivella.[72] data, the MeRGE datasets[92 93], and the IMPACT database[74 76 79].

7.3.2 Description of the 6 articles and my analysis of IPD available.

The IPD and summary data obtained related to six articles; Trivella et al.[72], the two MeRGE datasets[92 93], and three IMPACT articles[74 76 79]. These are summarized in Table 7.1 which shows for each IMPF article the PF and outcome, whether there is potentially availability bias, whether the authors originally examine small-study effects or publication bias, and what methods are used to analyse the IPD. None of these 6 articles had considered small-study effects previously.

The IPD for the IMPACT dataset contained all PFs and the outcomes they considered; PF values with and without missing data imputed, and an extended set of 15 studies (compared to studies in their original analyses) were available; so there was a wide range of potential analyses I could undertake. I decided to focus on the PF variables considered in McHugh et al.[79], Beek et al.[76], and Murray et al.[74] in relation to 6 month mortality. Further, I considered small-study effects in the original set of 7 studies and then again in the update set of 15 studies.

In Trivella et al.[72] the authors investigate whether microvessel density count after adjusting by age and stage is still a PF. Two sets of data available (all vessel method and Chalkely method), as there are two different methods of measurement. So I assessed small-study effects for each group.

In MeRGE1[92], the authors investigate whether there is an association between Restrictive Filling Pattern (RFP) and mortality rate for Acute Myocardial infarction disease; in particular they assess whether RFP is a PF before and after adjusting by Killip I, II and III/V; so I examined the small-study effects for unadjusted and adjusted results.

In MeRGE2[93], the authors examine whether there is an association between RFP and mortality rate for the patients with heart failure and whether RFP is a PF after adjusting by LV ejection fraction (LVEF) and age. Thus, again I examined small-study effects for both unadjusted and adjusted results.

7.3.3 My strategy for examining small-study effects

Following my experience of analysis the Look et al.[14] data (section 7.4), I decided upon a strategy in the 6 articles.

1. Assess small-study effects visually by using ‘contour funnel plot’ and ‘cumulative forest plot’.
2. Estimate the number of missing studies by using fixed random trim and fill method.
3. Obtain P-value for Egger’s test of asymmetry.

Table 7. 1: The brief summary for the 6 IMPF articles for which I could assess small-study effects

Study	PF of interest	Disease area	Outcome	No. of studies available to me	Percentage of IPD obtained	Was there potentially availability bias (i.e. studies for which IPD not obtained originally)	Stated the methods I used to obtain PF estimate and its s.e. In each study	Did the original article exam the small-study effects
Trivella et al[72].	Microvessel-density counts	Non-metastatic surgically treated non-small-cell lung carcinoma	Overall Survival	13	34%	Yes, because they did not provide all IPD studies that they asked for	Cox PH model	No, but they included 2 unpublished studies
MeRGE 1[92]	Restrictive Filling Pattern (RFP) was the key interest; they also want to adjust for the other factors such as Killip class	Acute Myocardial infraction	Mortality	18	56%	Yes, because they did not provide all IPD studies that they asked for	Cox PH model	No
MeRGE2[93]	Restrictive Filling Pattern (RFP) was the key interest; and they also adjusted by LVEF and age	Heart Failure	Mortality	6	35%	Yes, because they did not provide all IPD studies that they asked for	Cox PH model	No
Beek et al.[76]	Laboratory parameter (glucose, Prothrombin (Ph), haemoglobin (Hb))	Traumatic brain injury	Six month Glasgow outcome scale	15	NA	Not relevant, as it is collaboration for 11 studies.	logistic regression *	No
Mchugh et al.[79]	Secondary insult (e.g. hypoxia, hypotension, and hyotheria)	Traumatic brain injury	Six month Glasgow outcome scale	15	NA	Not relevant, as it is collaboration for 11 studies.	logistic regression*	No
Murray el al.[74]	Age, Glasgow coma scale (GCS), motor scale, pupil response, computerized tomography (CT), prothrombin time, hypotension, hypoxia, and glucose	Traumatic brain injury	Six month Glasgow outcome scale	15	NA	Not relevant, as it is collaboration for 11 studies.	logistic regression*	No

*All continuous PFs were analysed on their continuous scale and a linear trend assumed for simplicity. NA refers to not applicable

Then, if trim and fill method yielded 2 or more missing studies, or Egger's P-value was less than 0.10, or if visually I was concerned about small-study effects, I proceeded to

4. Use Egger's D-var method to estimate the adjusted pooled effect size of the PF.

Note that (2) to (4) were only considered if the number of studies was greater than 10.

7.3.4 The assessment of the small-study effects in the six IMPF articles

In this section, I assess small-study effects in the 6 IMPF articles outlined in section 7.3.2. First, I assess the three articles that used a literature reviews to identify relevant studies (Trivella et al.[14], MeRGE1[92] and MeRGE2[93]). Recall that literature review articles sought IPD by first identifying relevant studies using a literature review(e.g. using Medline, Embase), and then contacting authors of all identified articles to ask for their IPD (Chapter 2). In contrast, the three IMPACT articles (Beek et al.[76], Mchugh et al.[79], and Murray et al.[74]) use IPD from an existing dataset. Table 7.2 and 7.3 summarize the results found using my strategy outlined in 7.3.3.

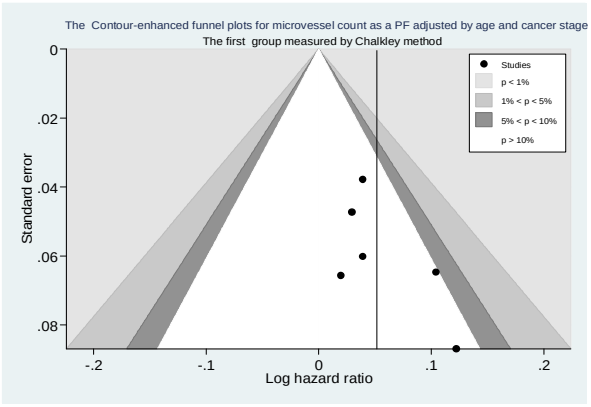
7.3.4.1 The assessment of the small-study effects for the Literature review articles

For Trivella et al.[72] I assess small-study effects in the analysis of whether microvessel count in a PF after adjusting by age and cancer stage, first for all vessel method and then Chalkely method. Figure 7.4 shows the contour funnel plots and cumulative forest plots for the two- measurement methods. The contour plot, for all vessel method with 13 studies, shows no evidence of asymmetry (Figure 7.4). In Chalkely method, there are only small differences between the standard error values across the 6 studies, and so it is not appropriate or sensible to examine small-study effects for this method as no smaller study exist[160]. For all vessel method the cumulative forest plots show barely any shift as smaller studies are included (Figure 7.4). There are no missing studies estimated by using trim and fill method, and Egger's P-values were greater than 0.10 (see Table 7.2). Thus, I conclude there is no concern for the small-study effects

for either measurement method, and this strengthens the original conclusions from Trivella et al.[72]

Figure 7.4: Funnel and cumulative funnel meta-analysis plot based on the log hazard ratio and its standard error for the first and second group association between risk of death and increase of microvessel count, as measured by chalkely and all vessel methods respectively adjusted for age and cancer stage in Trivella et al.[72] study.

Figure (7.4a): Contour funnel plot for microvessel count adjusted by age and cancer stage measured by all Chalkely method –for 6 studies



Figure(7.4 b): Cumulative funnel plot for microvessel count adjusted by age and cancer stage measured by Chalkely method –for 13 studies

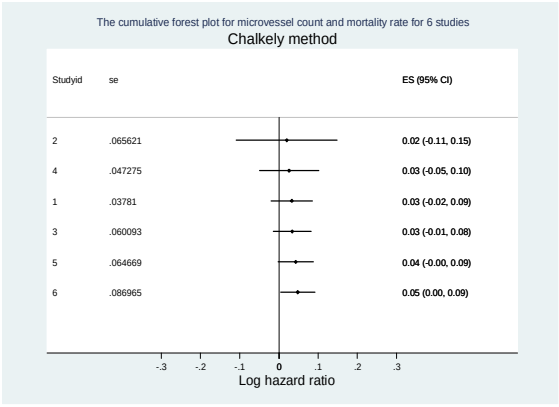


Figure (7.4c): Contour funnel plot for microvessel count adjusted by age and cancer stage measured by all vessel method –for 13studies

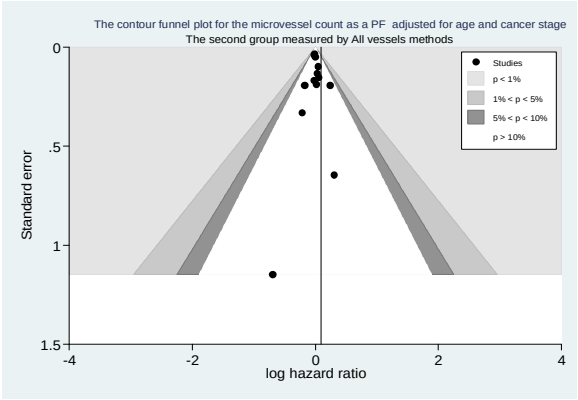
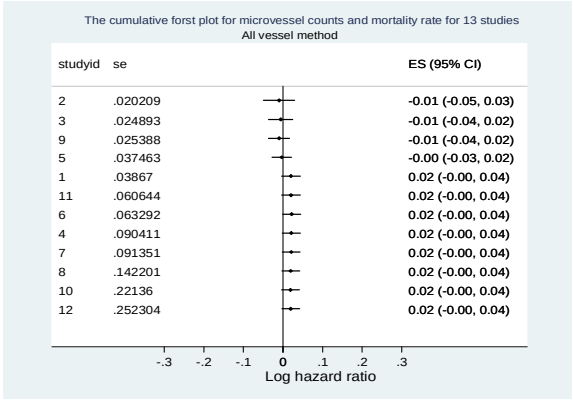


Figure (7.4d): Cumulative funnel plot for microvessel count adjusted by age and cancer stage measured by all vessel method –for 13 studies



Blue line is the meta-analysis pooled result.

In MeRGE1[92], I assessed the small-study effects for unadjusted RFP results and then after adjusting by LVEF and age. Figure 7.5 shows that for unadjusted and adjusted RFP, the contour funnel plots show no clear asymmetry and cumulative forest plots show no considerable shift as smaller studies are included. Trim and fill methods yielded 2 missing studies for the adjusted RFP, and Egger's P-value was 0.49. Based on my strategy in section 7.3, I was now only concerned about small-study effects in the analysis of adjusted results. The pooled effect size of

the original meta-analysis of the PF was 2.28 with 95% confidence interval (1.90 to 2.09), this indicates that RFP is a PF even after adjusting by LVEF and age; the adjusted pooled effect size from D-var method was 1.90 with 95% confidence interval (1.06 to 2.09), this indicates that although there is slight concern for small study effects, RFP (after adjusted by LVEF and age) still appear to be a PF even when adjusting for any funnel plot asymmetry. This again strengthens and enhances the original conclusion that RFP is prognostic.

Figure 7.5: The contour funnel plot for the log hazard ratio of RFP itself and after adjusted by LVEF and age, for the MeRGE1 study[92]

Figure 7.5a: The contour funnel plot for the log hazard ratio of unadjusted RFP, for18 IPD studies in MeRGE1

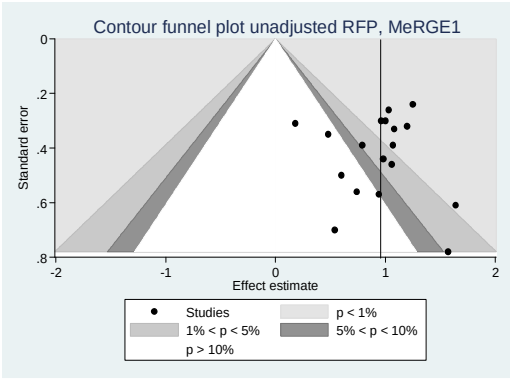


Figure 7.5b: The contour funnel plot for the log hazard ratio of unadjusted RFP for 18 IPD studies in MeRGE1

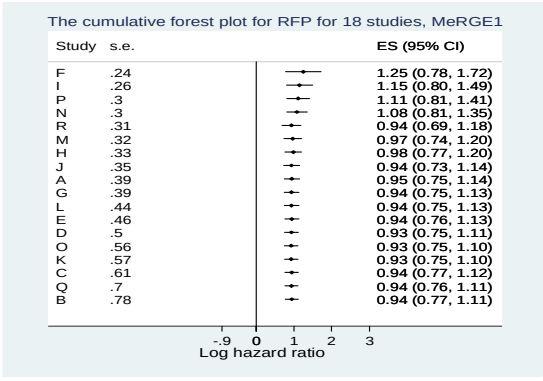


Figure 7.5c: The contour funnel plot for the log hazard ratio of RFP after adjusting by LVEF and age, for18 IPD studies in MeRGE1

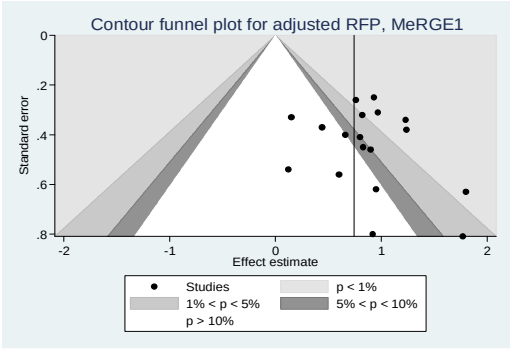
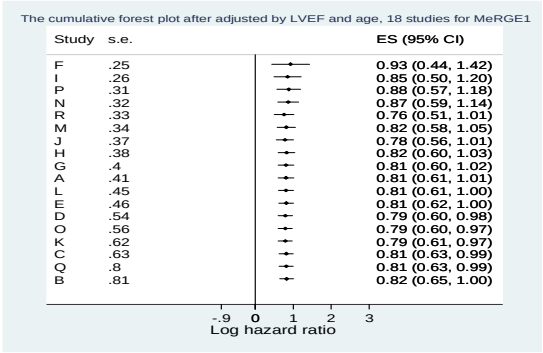


Figure 7.5d: The cumulative forest plot for the Hazard ratio of RFP after adjusting by LVEF and age , for18 IPD studies in MeRGE1



Black line is the meta-analysis pooled result.

Table 7. 2 : The assessment for the small-study effects and publication bias for the Trivella, MeRGE1 and MeRGE2, by using contour funnel plot, cumulative meta-analysis, Egger's methods and trim and fill method

Article	PF	Number of studies	Visual asymmetry	Cumulative meta-analysis shift	Egger's P-value	Trim and fill, number of missing studies	Original meta-analysis, with 95% CI	Original Egger's method, with 95% CI	Egger's D-var method, with 95% CI	Overall concerned about small-study effects
Trivella et al. (Hazard ratio)	Microvessel count adjusted for age , and cancer stage for All vessel method	13	No	Minor shift	0.68	0	------	------	------	No concern*
	Microvessel count adjusted for age , and cancer stage, for Chalkely method	6	No	Minor shift	------	------	------	------	------	No concern*
MeRGE1 Hazard ratio	RFP, whole dataset With 18 IPD studies	18	Tentative yes	Minor shift	0.90	1	------	------	------	No concern
	RFP after adjusted by LVEF and age , With 18 IPD studies	18	Tentative yes	Slight shift	0.49	2	2.28 (1.91 to 2.73)	2.06 (1.53 to 2.77)	1.90 (1.06 to 3.40)	There is a slight concern*, but conclusion does not change
MeRGE2 Hazard ratio (Hazard ratio)	RFP	6	No	Minor shift	------	------	------	------	------	No concern*
	RFP With info on Killip class	6	No	Minor shift	------	------	------	------	------	No concern*
	Killip Class II	6	No	Minor shift	------	------	------	------	------	No concern*
	Killip III/IV	5	Tentative yes	Minor shift	------	------	------	------	------	No concern*

*The number of the studies is not enough as it is less than 10 studies, e.g. this causes Egger D-var to have very wide confidence interval.

Figure 7.6: The cumulative forest plot and contour funnel plot for RFP itself and after adjusting by killip and grouped by killip II and III/IV, for MeRGE2 study

Figure 7.6a : The cumulative forest plot for RFP itself

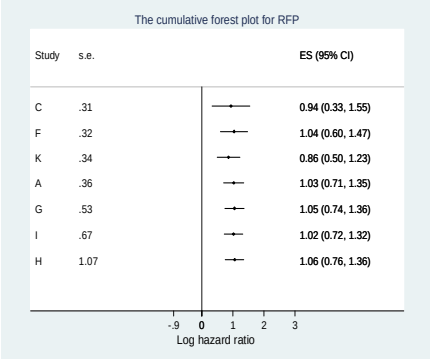


Figure 7.6b: The cumulative forest plot for RFP after adjusted by Killip

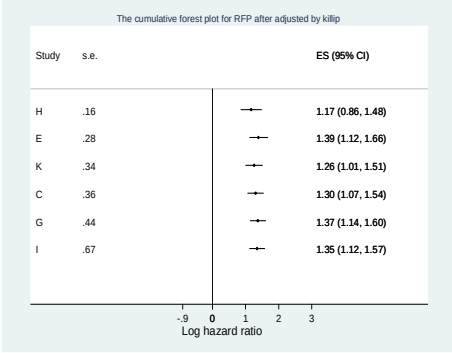


Figure 7.6c: The cumulative forest plot for RFP for Killip II

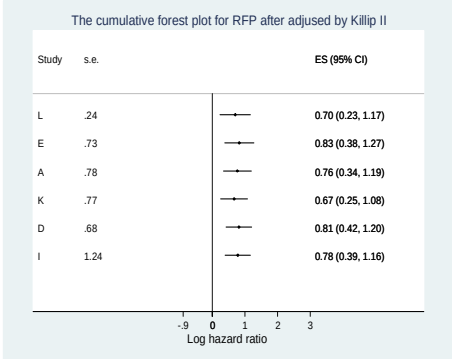


Figure 7.6d: The cumulative forest plot for RFP for Killip III/IV

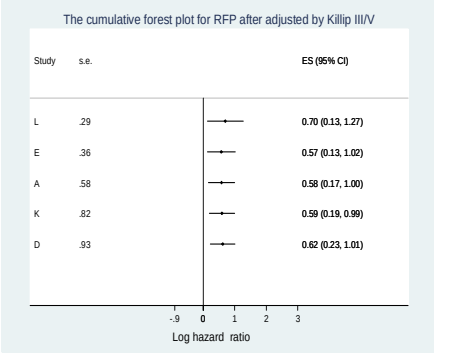


Figure 7.6e : The contour funnel plot for RFP itself

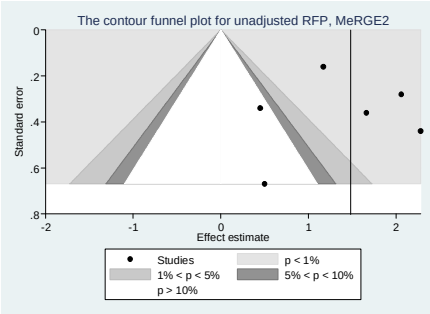


Figure 7.6f : The contour funnel plot for RFP after adjusted by Killip

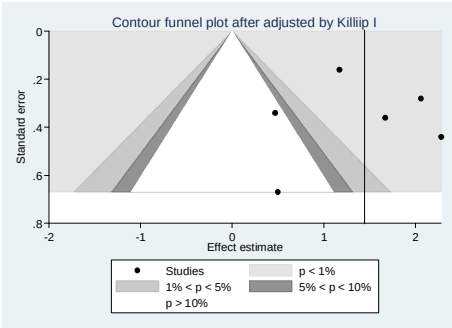


Figure 7.6g : The contour funnel plot for RFP for Killip II

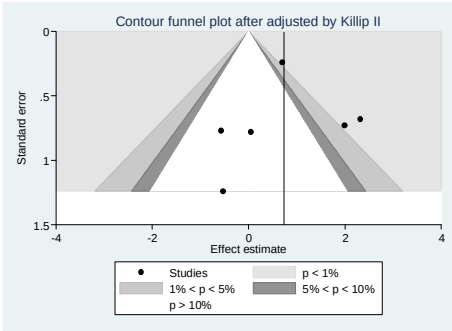
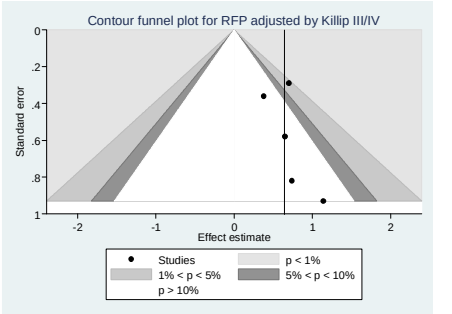


Figure 7.6h : The contour funnel plot for RFP for Killip III/IV



Black line is the meta-analysis pooled result.

In MeRGE2[93], I assessed the small-study effects for RFP as an independent PF, and after adjusting by Killip, and grouped by Killip II, and III/IV. Note that as the number of studies is less than 10 studies, I do not apply trim and fill method or Egger's test etc. Visually, there is no concern for small-study effects (Figure 7.6).

7.3.4.2 The assessment of the IMPACT articles

Table 7.3 summarizes my investigation of small-study effects for the three IMPACT studies. Figures 7.7 to 7.12 show the contour funnel plots and cumulative forest plots for the three IMPACT studies. Generally, there is no evidence that small study effects are a concern. Even when Egger P-value is less than 0.10 or trim and fill suggest greater than two-missing studies, the asymmetry is in the opposite direction to a publication bias mechanism. For example, in i_hyypoxia results for 14 IPD studies for Mchugh the original meta-analysis of the PF was 1.76 with 95% confidence interval 1.45 to 2.14, whilst using Egger's D-var adjust for asymmetry gives an adjusted pooled effect size of i_hypoxia was 2.28 with 95% confidence interval 1.94 to 2.67. The only slight cause for concern is for age with original 11 studies, as there is asymmetry with missing studies potentially in the white area of the graph closer to the null value, which may suggest the problem of publication bias. Also, there is a gradual shift to the right hand side of the cumulative forest plot. Trim and fill method yielded 5 missing studies. Whilst the original meta-analysis results of the PF was 2.09 with 95% confidence interval (1.79 to 2.45), the adjusted pooled effect size estimate from Egger's D-var method was 1.82 with 95% confidence interval between (1.61 and 2.06). However, this suggests age is still a PF and so-despite the asymmetry-the conclusion does not change.

To sum up, for all variables, there is barely any concern for small-study effects and there is no evidence that small-study effects would change the conclusion of the original data, in other words, the variables of interest are all PFs and the investigation of small-study effects has strengthened this conclusion.

Figure 7.7: The contour funnel plot for the original IPD studies that included in the original article and the 15 recent IPD studies for Beek et al.[76]

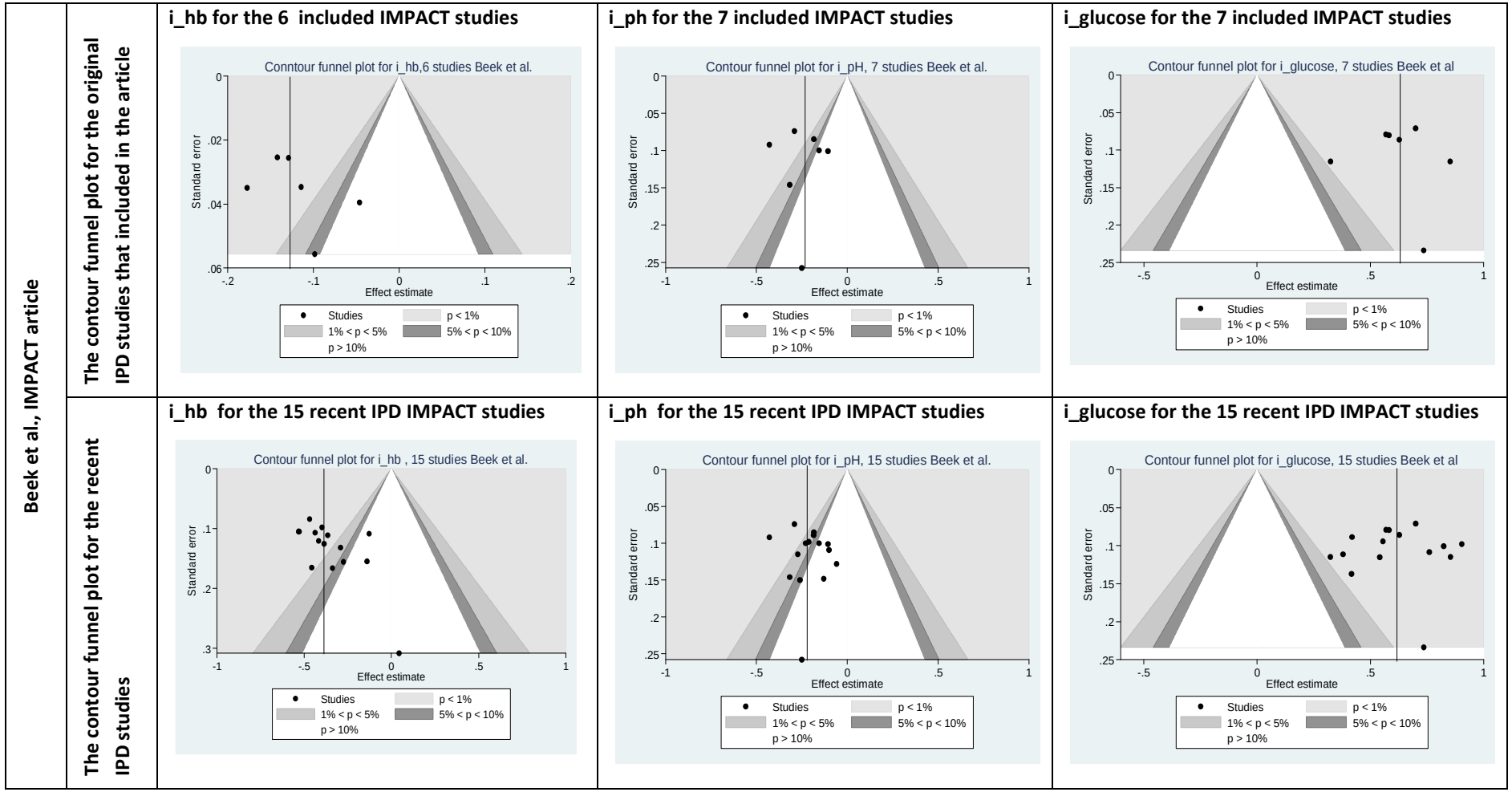


Figure 7. 8: The cumulative funnel plot for the original IPD studies that included in the original article and the 15 recent IPD studies for Beek et al.[76]

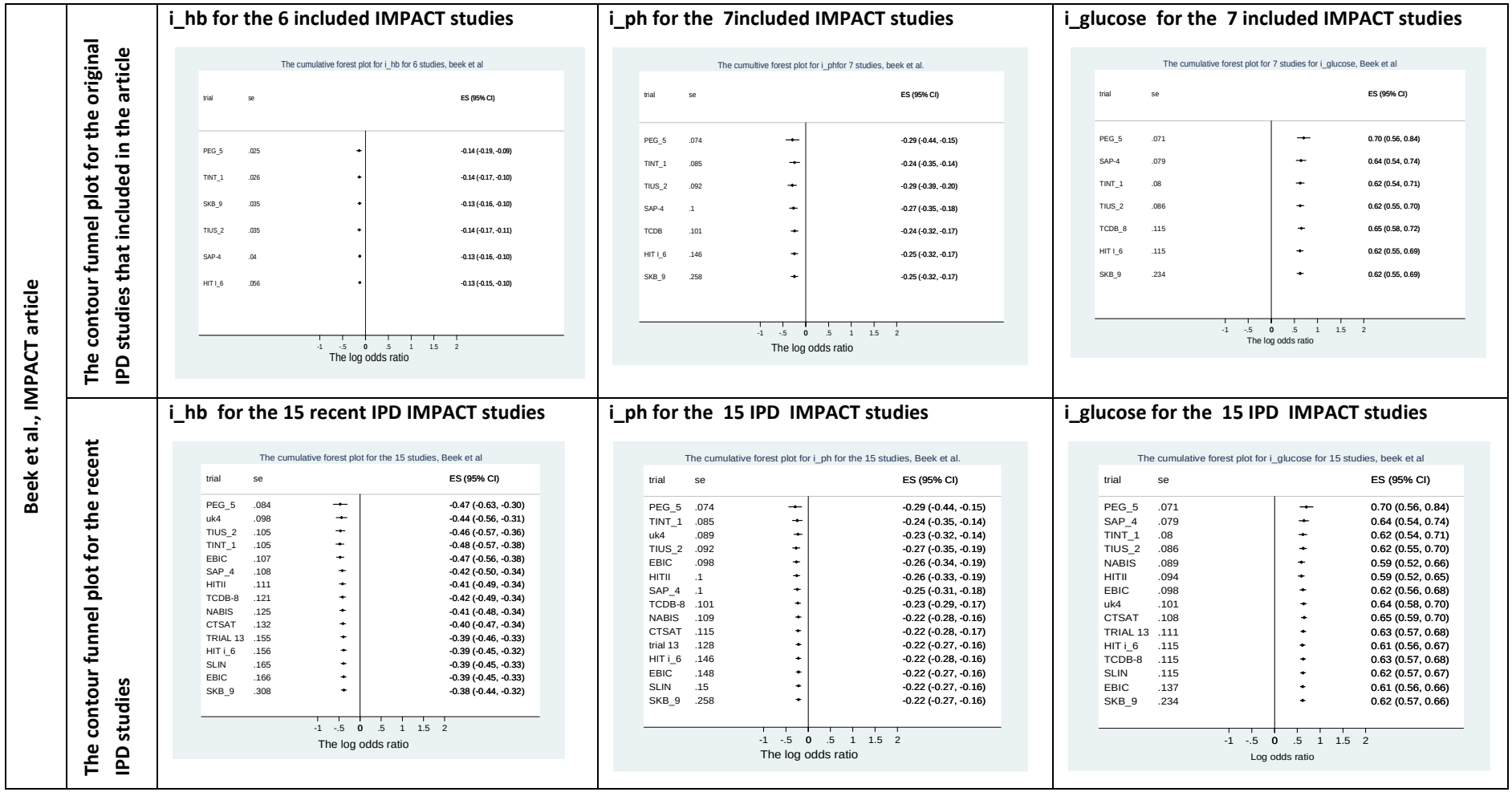


Table 7. 3 : The assessment for the small-study effects and publication bias for the three IMPACT studies (Beek et al.[76], McHugh et al.[79] and Murray et al.[74]), by using contour funnel plot, cumulative meta-analysis, Egger's methods and trim and fill method

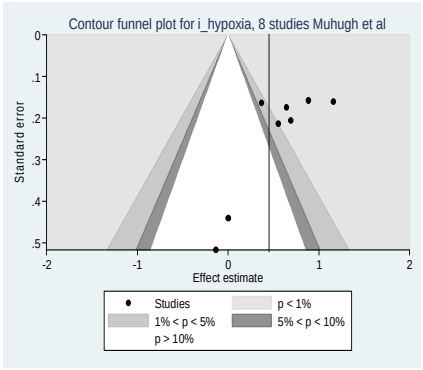
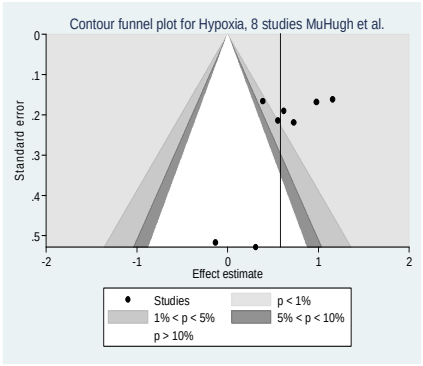
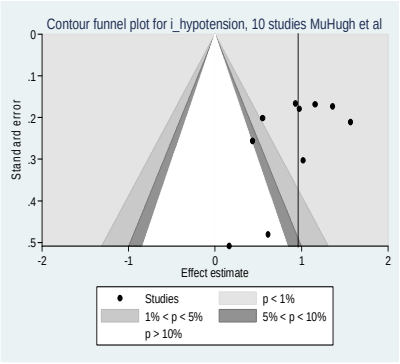
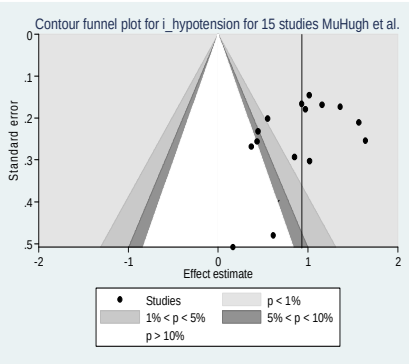
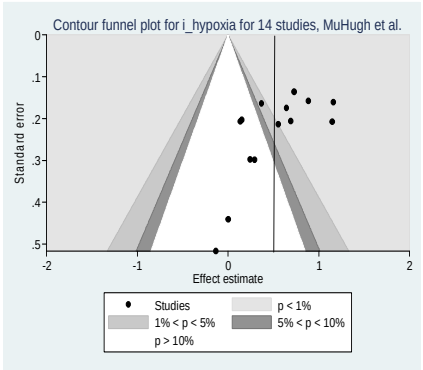
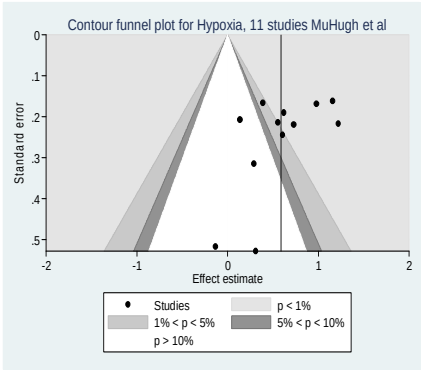
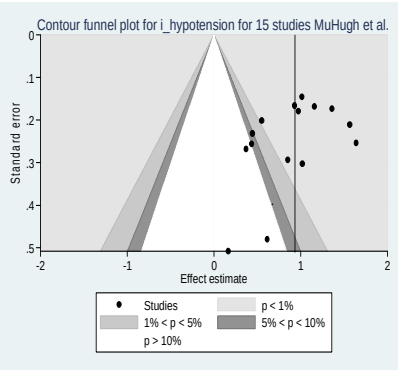
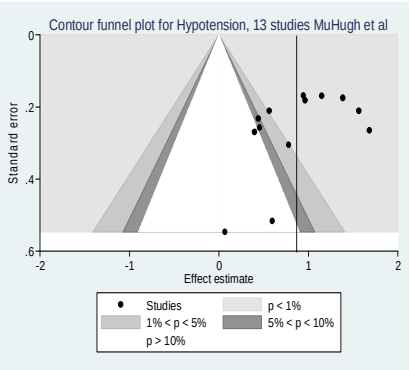
Article	Analysis	PF	No. of studies	Visual asymmetry	Cumulative meta-analysis shift	Egger's P-value	Trim and fill, number of missing studies	Original meta-analysis, with 95% CI	Egger's D-var method, with 95% CI	Overall concerned about small-study effects
Beek et al.	Original IPD studies	i_hb for 6 IPD studies	6	Tentative yes	No shift	------	------	------	------	slight concern
	The recent IPD studies	i_hb for the recent 15 studies	15	Tentative yes	Slight right shift	0.05	4	0.68 (0.64 to 0.72)	0.62 (0.56 to 0.69)	Slight concern
	Original IPD studies	i_ph for 7 IPD studies	7	Tentative yes	No shift	------	------	------	------	No concern
	The recent IPD studies	i_ph for 15 IPD studies	15	No	Slight right shift	0.56	2	0.80 (0.76 to 0.85)	0.79 (0.72 to 0.87)	Slight concern
	Original IPD studies	i_glucose for 7 IPD studies	7	Tentative yes	Slight left shift	------	------	------	------	No concern
	The recent IPD studies	i-glucose for 15 IPD studies	15	No	Slight left shift	0.53	0	------	------	No concern
Mchugh et al., IMPACT article. Odds ratio	Original IPD studies	i_Hypoxia for the 8IPD studies	8	Tentative yes	Slight left shift	0.12	0	------	------	No concern
	The recent IPD studies	i_Hypoxia for the 14 IPD studies	14	Tentative yes	Slight shift to the left	0.06	0	1.76 (1.45 to 2.14)	2.28 (1.94 to 2.67)	Slight concern
	Original IPD studies	Hypoxia for the 8IPD studies	8	Tentative yes	Slight left shift	0.21	0	------	------	No concern
	The recent IPD studies	Hypoxia for the 14 IPD studies	11	Tentative yes	Slight left shift	0.26	0	------	------	No concern
	Original IPD studies	i_Hypotension for the 10 IPD studies	10	Tentative yes	Slight shift to the left	0.21	0	------	------	No concern
	The recent IPD studies	i_Hypotension for the 14 IPD studies	14	Tentative yes	Slight shift to the left	0.20	0	------	------	No concern
	Original IPD studies	i_Hypotension for the 9 IPD studies	9	Tentative yes	No shift	------	------	------	------	No concern
	The recent IPD studies	i_Hypotension for the 13 IPD studies	13	Yes	No shift	0.18	0	------	------	No concern

Note that i_hb: is the haemoglobin with imputed the missing data, i_pH is the Prothrombin with imputed the missing data, i_glucose: is the glucose after imputed the missing data, etc.

Article	Analysis	PF	No. of studies	Visual asymmetry	Cumulative meta-analysis shift	Egger's P-value	Trim and fill, number of missing studies	Original meta-analysis, with 95% CI	Egger's D-var method, with 95% CI	Overall concerned about small-study effects
Murray et al.	Original IPD studies	Pupil for the 11 IPD studies	11	Tentative yes	Slight right shift	0.94	0	------	------	No concern
	The recent IPD studies	Pupil for the 14 IPD studies	15	Tentative yes	Slight right shift	0.88	0	------	------	No concern
	Original IPD studies	Age for 11 studies	11	Tentative yes	Slight right shift	0.21	5	2.09 (1.79 to 2.45)	1.82 (1.38 to 2.44)	Slight concern
	The recent IPD studies	Age for 15 studies	15	Tentative yes	Slight right shift	0.19	0	------	------	No concern
	Original IPD studies	Adjusted Hypoxia for 11 studies*	11	Tentative yes	Slight right shift	0.09	0	1.57 (1.32 to1.85)	1.86 (1.46 to 2.36)	Slight concern
	The recent IPD studies	Adjusted hypoxia for the 15 studies*	15	Yes	Slight right shift	0.07	0	1.52 (1.32 to 1.76)	1.86 (1.46 to 2.29)	Slight concern
	Original IPD studies	Adjusted Hypotension, for 11 studies*	11	Yes	Slight right shift	0.22	0	------	------	No concern
	The recent IPD studies	Adjusted Hypotension, for 15 studies*	15	Yes	Slight right shift	0.16	0	------	------	No concern
	Original IPD studies	Adjusted i_glucose, for 11 studies *	11	Tentative yes	Slight right shift	0.96	1	------	------	No concern
	The recent IPD studies	Adjusted i_glucose , for 15 studies*	15	Yes	Slight right shift	0.71	0	------	------	No concern
	Original IPD studies	Adjusted pH, for 11 studies*	11	Tentative yes	Slight right shift	0.14	1	------	------	No concern
	The recent IPD studies	Adjusted pH, for 15 studies*	15	Yes	Slight right shift	0.02	5	0.73 (0.64 to0.78)	0.66 (0.59 to 0.74)	Slight concern

* Note that Hypoxia, Hypotension, i_glucose, and pH are adjusting by age d_motor, and pupils, for original studies included and the 15 recent IPD studies

Figure 7.9: The contour funnel plot for i_hypoxia, hypoxia, hypotension and i_hypotension for the included studies in the original Mchugh et al.[79], and the recent IPD IMPACT studies.

Mchugh et al., IMPACT article.	**The contour funnel plot for the original IPD studies that included in the article**	**i_Hypoxia for the 8 included IMPACT studies** 	**Hypoxia for the 8 included IMPACT studies** 	**i_Hypotension for the 10 included IMPACT studies** 	**Hypotension for the included IMPACT studies** 
	The contour funnel plot for the recent IPD studies	**i_Hypoxia for the 14 recent IMPACT IPD studies** 	**Hypoxia for the 11 included IMPACT studies** 	**i_Hypotension for the recent IPD IMPACT studies** 	**Hypotension for the recent IPD IMPACT studies** 

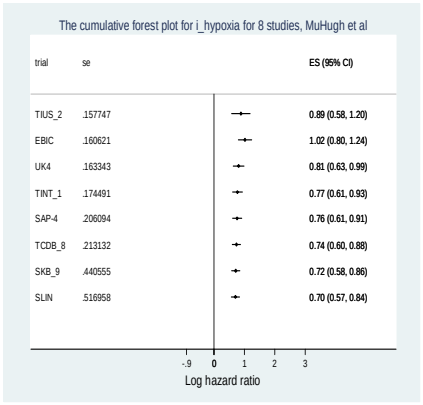
Black line is the meta-analysis pooled result.

Figure 7.10: The cumulative forest plot for i_hypoxia, hypoxia, hypotension and i_hypotension for the included studies in the original Mchugh et al.[79], and the recent IPD IMPACT studies.

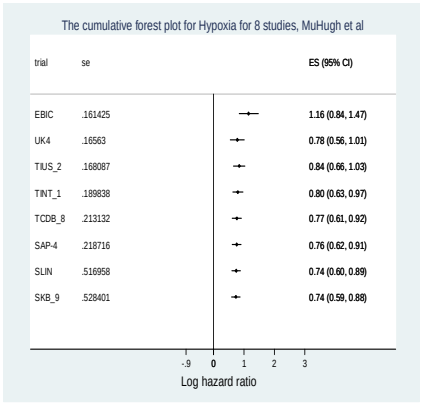
Mchugh et al., IMPACT article.

The cumulative forest plot for the original IPD studies that included in the article

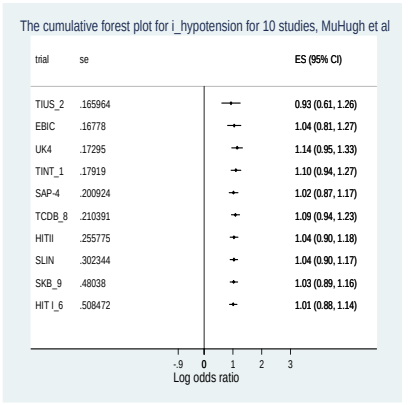
i_Hypoxia for the 8 included IMPACT studies



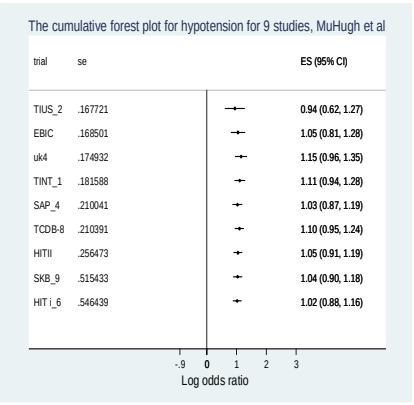
Hypoxia for the 8 included IMPACT studies



i_Hypotension for the 10 included IMPACT studies

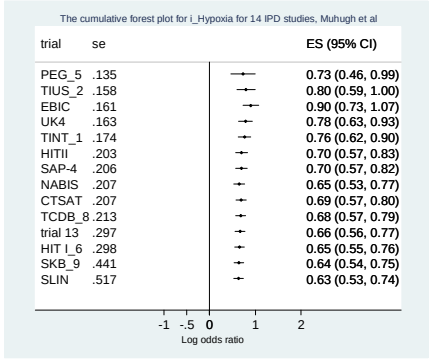


Hypotension for the 9 included IMPACT studies

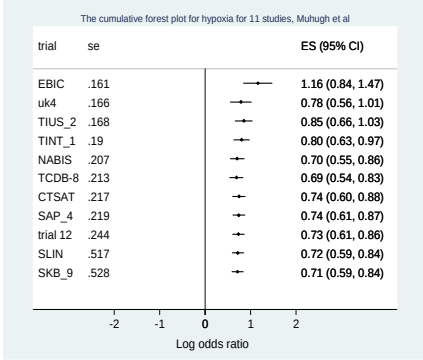


The cumulative forest plot for the recent IPD studies

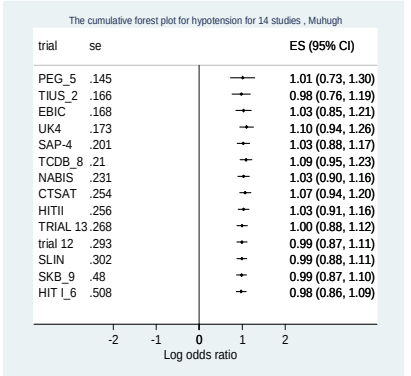
i_Hypoxia for the 14 recent IMPACT IPD studies



Hypoxia for the 11 included IMPACT studies



i_Hypotension for the recent IPD IMPACT studies



Hypotension for the recent IPD IMPACT studies

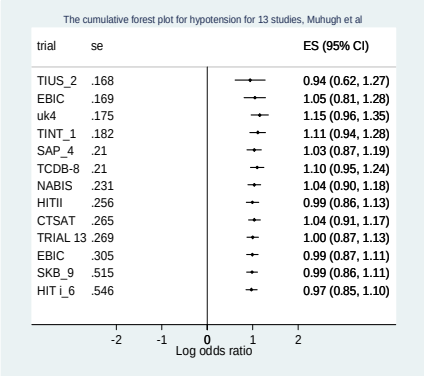
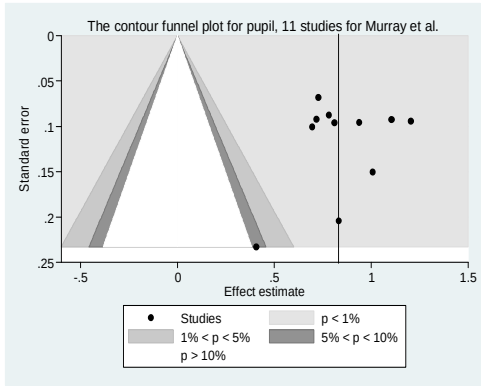
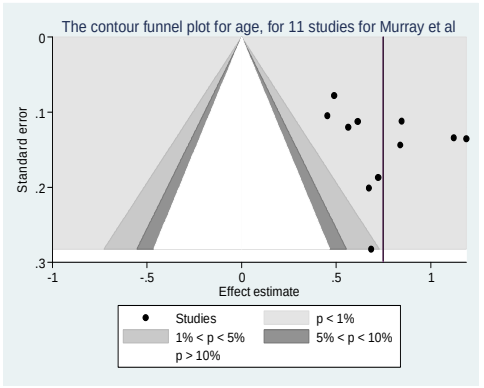
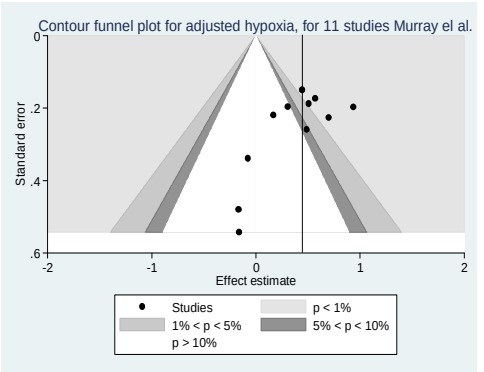


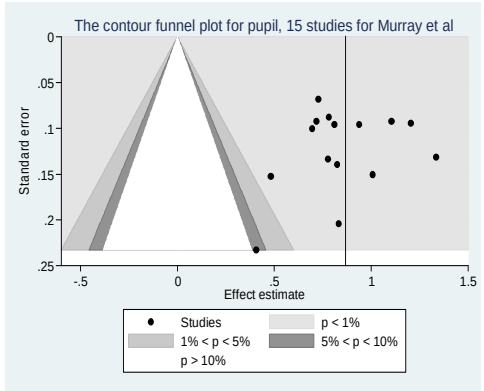
Figure 7.11: The contour funnel plot for pupil, age, hypoxia, hypotension, i_glucose and pH after adjusted by age, d_motor, and pupils for the included studies in the original Murray et al.[74], and the recent IPD IMPACT studies.

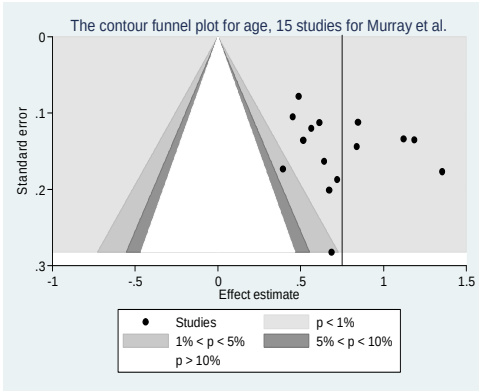
Murray et al., IMPACT article	**The contour funnel plot for the included IPD studies**	**Pupil for the 11 included IPD of IMPACT studies**	**Age for the 11 included IPD of IMPACT studies**	**hypoxia after adjusted by age d_motor, and pupils for the included IPD IMPACT studies**
	The contour funnel plot for the recent IPD studies	**Pupil for the 15 recent IPD IMPACT studies**	**Age for the 15 recent IPD IMPACT studies**	**hypoxia after adjusted by age d_motor, and pupils for the recent IPD IMPACT studies**

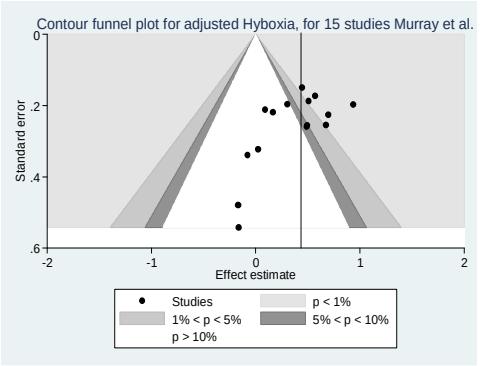








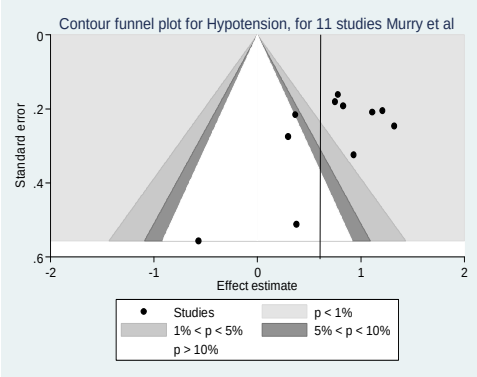
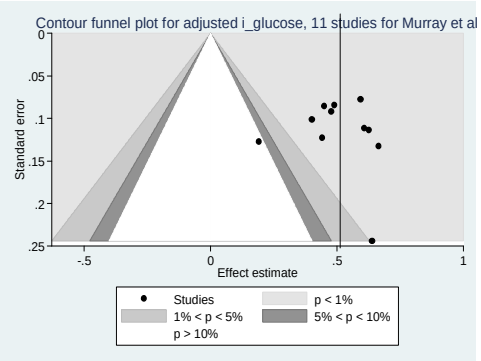
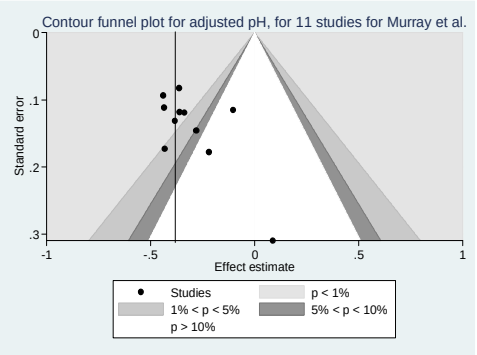
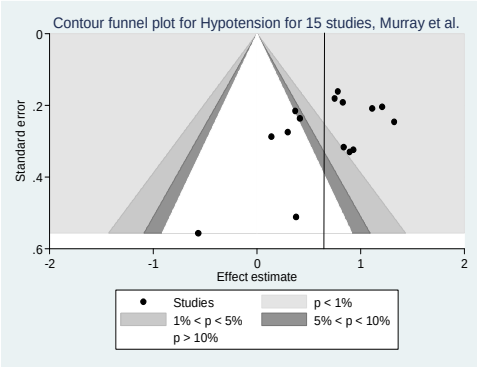
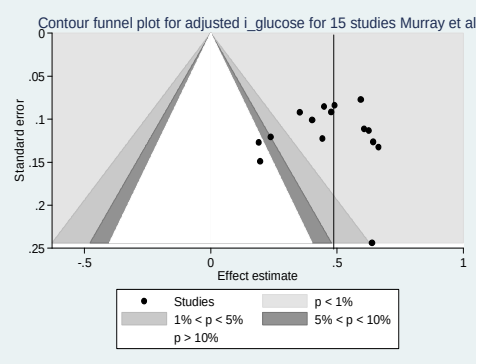
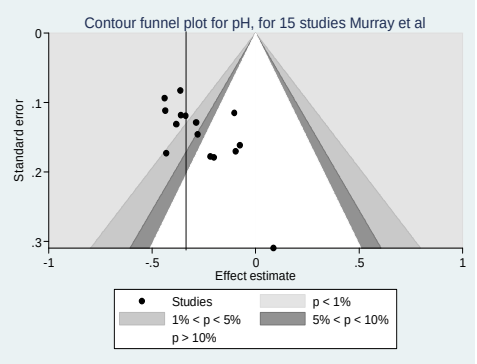




The black line is the original meta-analysis results

Continue on the next page

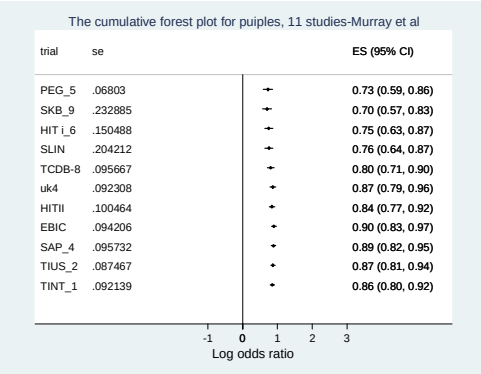
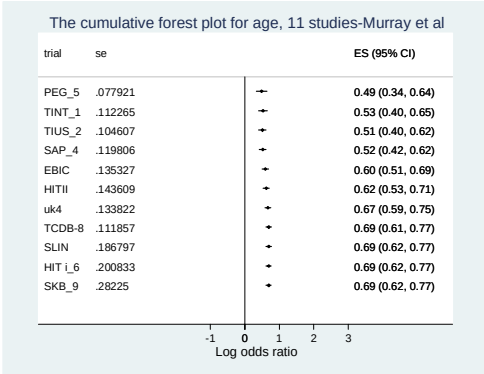
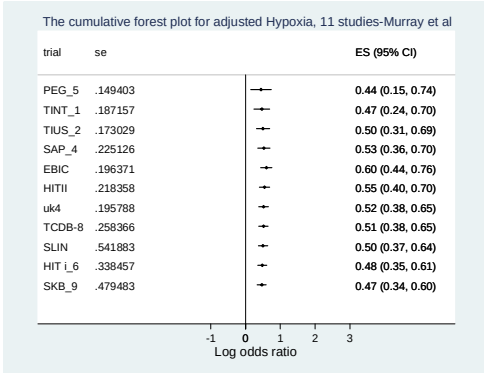
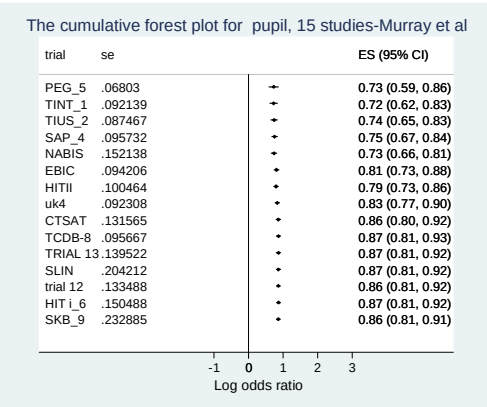
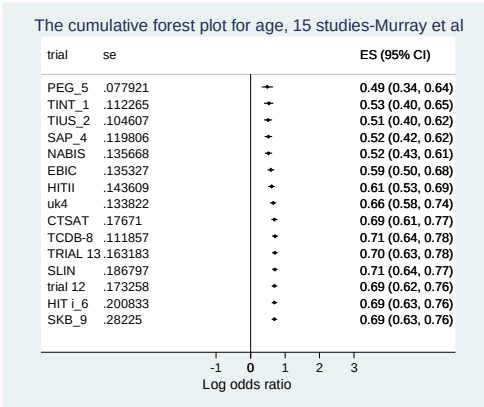
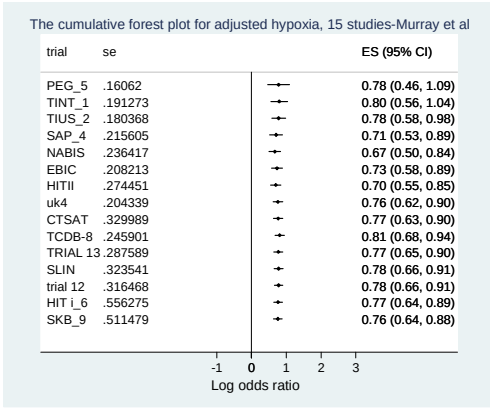
Continue from previous page

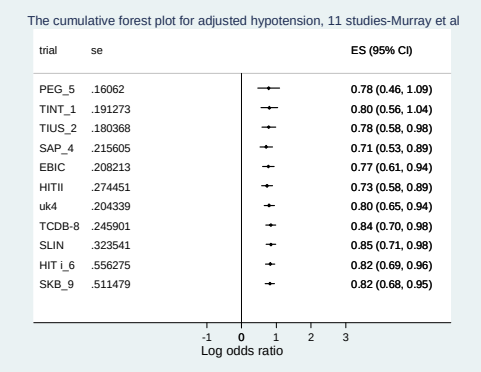
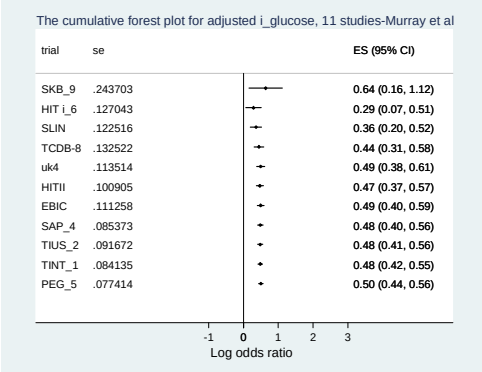
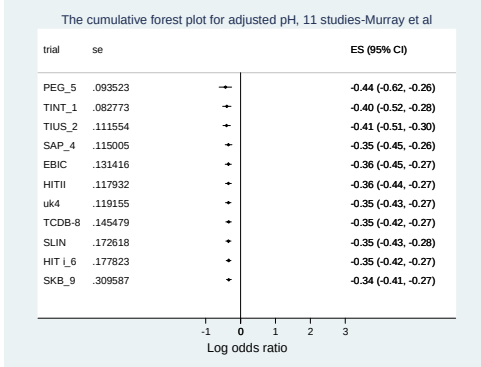
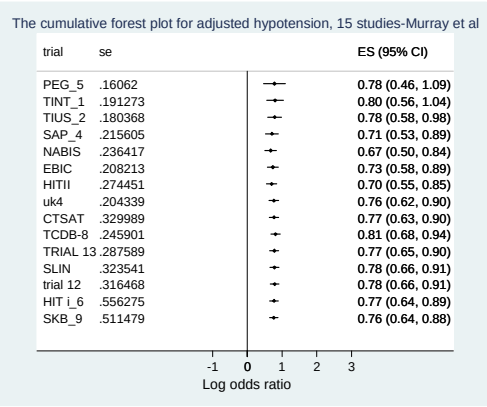
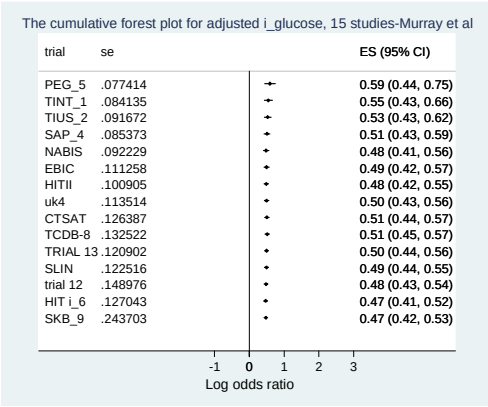
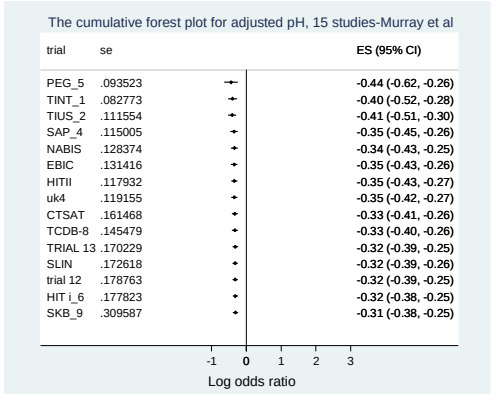
Murray et al., IMPACT article	**The contour funnel plot for the included IPD studies**	**Hypotension after adjusted by age d_motor, and pupils for the included IMPACT studies** 	**i_glucose after adjusted by age d_motor, and pupils for the included IPD IMPACT studies** 	**pH after adjusted by age d_motor, and pupils for the included IPD IMPACT studies** 
	The contour funnel plot for the recent IPD studies	**Hypotension after adjusted by age d_motor, and pupils for the recent IPD IMPACT studies** 	**I_glucose after adjusted by age d_motor, and puipls for the recent IPD IMPACT studies** 	**pH after adjusted by age d_motor, and puiplsfor the recent IPD IMPACT studies** 

The black line is the original meta-analysis results

End of the Figure

Figure 7.12: The contour funnel plot for pupil, age, hypoxia, hypotension, i_glucose and pH after adjusted by age, d_motor, and pupils for the included studies in the original Murray et al.[74], and the recent IPD IMPACT studies.

<table>
<tr>
<td rowspan="2">Murray et al., IMPACT article.</td>
<td>The cumulative forest plot for the included IPD studies</td>
<td>Pupil for the 11 included IMPACT studies
 
</td>
<td>Age for the 11 included IMPACT studies
 
</td>
<td>hypoxia after adjusted by age d_motor, and pupils for the included IPD IMPACT studies
 
</td>
</tr>
<tr>
<td>The cumulative forest plot for recent IMPACT IPD studies</td>
<td>Pupil for the 15 recent IPD IMPACT studies
 
</td>
<td>Age for the 15 IPD IMPACT studies
 
</td>
<td>hypoxia after adjusted by age d_motor, and pupils for the recent IPD IMPACT studies
 
</td>
</tr>
</table>

FMurray et al., IMPACT article	**The cumulative forest plot for the included IPD studies**	**Hypotension after adjusted by age d_motor, and puipls for the included IMPACT studies** 	**i_glucose after adjusted by age d_motor, and pupils for the included IPD IMPACT studies** 	**pH after adjusted by age d_motor, and pupils for the included IPD IMPACT studies** 
	The cumulative forest plot for recent IMPACT IPD studies	**Hypotension after adjusted by age d_motor, and pupils for the recent IPD IMPACT studies** 	**I_glucose after adjusted by age d_motor, and pupils for the recent IPD IMPACT studies** 	**pH after adjusted by age d_motor, and puipls for the recent IPD IMPACT studies** 

7.4 Conclusion

In this chapter, I undertook an assessment of publication bias for 6 articles[14 72 74 76 79 92 93]by using contour funnel plot, cumulative funnel plot, trim and fill method and Egger's test. I e-mailed the authors of 19 IMPF articles (as there is one article[72] reported their aggregated data in the article) asking them whether it is possible to provide the aggregated data (effect size of the PF and its standard error) or their IPD, but only 2 articles provide the aggregated data, and the 6 IMPACT articles provide the IPD data.

The results for these articles generally give a little evidence of small-study effects; this indicates that the IMPF articles may have less concern for small-study effects or publication bias than meta-analysis of PFs using aggregated data.

The limitation here is that the assessment of small-study effects has been undertaken for only a sample IMPF projects that actually gave me data. Thus, my research could also be affected by bias; that study results were not obtained from most IMPF articles is very poor practice, and shows reporting of IPD meta-analysis most improve. Also, for simplicity I assumed linear trends for all continuous PFs. I also recognise that assessment of funnel plot asymmetry is notoriously subjective, and others may interpret results differently. However, I have used a clear strategy that only if there was clear visual asymmetry, a significant Egger's test results, or two or more missing studies estimated by trim and fill method would I be concerned about small-study effects. Only then did I use Egger's D-var regression method to obtain adjusted pooled estimates.

I undertook the empirical study to assess the small-study effects and publication bias in the IMPF articles. However, there are other empirical studies that assess the small-study effects in meta-analysis and clinical trials. For example, Palma et al.[168] undertook a systematic review of 225 meta-analyses, to examine the assessment for publication bias with meta-analysis on cardiovascular diseases. They stated that '*The frequency of assessment of*

publication bias in meta-analysis is still very low, although it has improved with time. It is more frequent in meta-analyses on observational studies and it is related to other methodological characteristics of reviews'. Sutton et al.[169] *took 48 reviews in Cochrane Database to assess the effect of publication bias on the results of these articles.* The authors stated *"Publication or related biases were common within the sample of meta-analyses assessed. In most cases these biases did not affect the conclusions. Nevertheless, researchers should check routinely whether conclusions of systematic reviews are robust to possible non-random selection mechanisms."*

There are also empirical studies to assess the small-study effects and publication bias for the prognosis research. For example, Kyzas et al.[170] found that most of the articles on cancer PF studies yielded significant results. This perhaps indicates to a major problem of publication bias, and Hemingway et al. undertook a systematic review for the prognostic effects of circulating biomarkers in stable coronary disease; they include 390 reports of biomarker effects, and state that "*The quality of individual study reports was variable, with evidence of small study (publication) bias and incomplete adjustment for simple clinical information such as age, sex, smoking, diabetes and obesity*". This highlights that there is an evidence for small-study effects in prognosis research.

In conclusion, although most of the 20 IMPF studies do not assess the small-study effects or publication bias in their analysis, after assessing whether there is evidence of small-study effects, I found that there is very little evidence of small-study effects to cause concern here (based on the 6 IMPF studies). Even, in Look et al. where small-study effects were found, adjusting for it did not change clinical conclusion, although the PF effect was smaller. I recommend all IMPF projects should assess small-study effects and potential publication bias in the future , as Hemingway et al.[2] state that **"***A prudent default position would be to*

assume that prognosis research is seriously afflicted by publication bias, until there is evidence to the contrary".

CHAPTER 8

DISCUSSION AND CONCLUSION

8.1 Introduction

Prognostic factors (PFs) are patient characteristics (e.g. age, biomarker levels) that are associated with future clinical outcomes in patients with existing disease. PFs are important clinical tools because they help to identify patients with different risks of outcome (e.g. recurrence of disease) and thereby facilitate the most appropriate treatment strategies and aid patient counselling[3], they are also used in the design and analysis of trials, and act as confounding factors in observational studies[4]. Evidence-based results regarding PFs are therefore very important to both clinicians and their patients. However, primary PF studies have numerous problems such as poor reporting, analysis and design[5-7]. PF studies are subject to numerous biases, such as selective reporting[7] and "optimal" choice of cut-points[8]. These severely limit meta-analysis of PF studies using aggregated data, where summary of results are combined across PF studies. Individual Patient Data Meta-analysis of PF studies (IMPF) have been proposed as the gold-standard approach[9-11], because it utilises the raw patient data and thus does not rely on reported results.

In this thesis I have formally demonstrated the challenges and the pitfalls of IPD meta-analysis of prognostic factor (IMPF) studies through a literature review of IMPF articles (Chapter 2). I have also performed a comparison study between the aggregated data and IPD approach by using an empirical example in breast cancer to show up which approach is more flexible and more reliable for statistical analysis. I found that IPD approach is better than aggregated data (Chapter 3) because it allows to adjust the PF by other confounding factors, assess the interaction between two PFs as an additional PF, and examine PFs in subgroup of

patients. But it still has challenges such as missing data, different methods of measurement, and potentially publication bias.

I have also demonstrated a wide range of possible one-step and two-step for IPD meta-analysis of PFs (11 models developed); which differed according to whether they account for the clustering of patients across studies, whether they accounts for the residual variation for the same values of the PF, and whether they account for the correlation of parameter estimates (Chapter 4). TBI dataset was used as an applied example to show how the effect size of the PF and its standard error changed accordingly for the eleven one-step and two-step models. Surprisingly, these were little difference between the models. But I recognized this was just one dataset for which the baseline risk was remarkably similar across studies.

So in chapter 5 a simulation study was undertaken to assess three key models from the 11 possible one-step and two-step models; in particular the standard two-step model (which account for the clustering of patients across studies); the one-step model that ignores the clustering of patients across studies and assumes that the whole IPD comes from one study; and the one-step model that accounts for the clustering of patients across studies by using an indicator variable. The simulation study was undertaken for a single binary or continuous PF in relation to a binary outcome, and I assumed the PF effect was fixed across studies and only a simple univariate logistic regression model was fitted. I found that one-step meta-analysis that account for the clustering across studies is the best fitted model, as it consistently produces unbiased estimates with suitable coverage.

Previous chapters assumed linear trends for continuous PFs. In chapter 6 I demonstrated all of the first and second order possible IPD logistic regression fractional polynomial models that can be used, if the linearity assumption of the PF does not hold. Again the TBI dataset was used as an applied example to show and highlight the accuracy for each model according

to chi-square statistics and deviance; in particular to indicate to the best model that has less deviance. I found one-step FP model is the easiest to fit and that a quadratic trend between age and the logit risk of 6 month mortality was appropriate.

Finally, in chapter 7 I examined the problem of small-study effects in IMPF articles. I was able to consider - through obtaining IPD and extracting aggregated data- statistical methods for detecting small-study effects and their potential impact on IMPF results. I found that there is very little evidence of small-study effects in IMPF articles. Even when small-study effects were a concern (e.g. in the best cancer data), adjusting for funnel plot asymmetry did not alter clinical conclusion.

Now, in this concluding chapter I will summarize and discuss the most important findings and issues that have arisen during the thesis, and will particularly emphasise why the research undertaken benefits future evidence syntheses of PF studies. I will also consider the numerous challenges that remain unaddressed by the research in this thesis, and will identify the main further research priorities that have specifically developed out of the work I have presented.

8.2 Key Findings and Recommendations for my Thesis

8.2.1 Benefits of IPD over aggregate data for meta-analysis of PF studies.

Individual Patient Data (IPD) is considered as a gold standard method for producing reliable PF meta-analysis results[107]. I have demonstrated some of these advantages , which include fitting one-step models, modelling non-linear trends, reducing the threat of small-study effects, and adjustment for other variables[171]. It also allows an evaluation of combinations of PFs, which may produce more specific and accurate prognostic assessments (Chapter 3).

IPD meta-analysis of PF is more preferable to using aggregated data. In particular, there are several limitations for meta-analysis of aggregated data[57] such as the variables included in the multivariable model are different from study to study, which is likely to cause heterogeneity in the multivariate PF results across studies; there is no same strategy for the choice of cut-off level among studies (different cut-off points across studies); there are missing values for the aggregated data (e.g. hazard ratio) in some studies; it also limits the ability to check the assumptions of the statistical model in each study, as well as the ability to look at the PF after adjusting or grouping by other variables. In contrast, IPD meta-analysis has many advantages over aggregated data[56 99] and Figure 8.1 lists some findings of the benefit of IPD approach[60], However this does not mean that IPD meta-analysis is without its challenges and pitfalls. In section 8.2.2 I tackled some of these challenges, and I highlighted some challenges that need to be considered in future research.

8.2.2 The Challenges Facing IMPF Projects

There are numerous methodological, statistical and practical problems facing IMPF projects (Chapter 2). Firstly, IPD does not solve the problem of poor primary studies. Riley et al.[22] stated that '*Prognostic markers are important tools in the management of patients with cancer and many other diseases, and as such primary studies of prognostic markers are essential. However, the design and evaluation of such studies can be greatly improved'*. Alongside, poor quality of primary studies there are challenges in statistical analysis such as how to tackle missing data and different methods of measurement; which method of meta-analysis is recommended (one-step or two-step); also whether accounting for the clustering of patients within studies is considered; how we handle continuous variables when the linearity assumption fails; publication bias and availability bias; checking the assumption of the statistical analysis in each study and what to do when it fails; also handling some other issues such as the problem of skewness in some continuous variables and dealing with outlier

values. In this thesis, I considered some of these challenges, but there are still some challenges need to be considered (See Chapter 2 Figure 2.8), and section 2.8.

Figure 8.1: Some advantages of using individual patient data for a meta-analysis

- Consistent inclusion and exclusion criteria can be used across studies, and, if appropriate, individuals who were originally excluded can be reinstated into the analysis
- Results presented in the original study publications will be verified (assuming individual patients data provided can be matched to the individual patients data used in the original analyses)
- Studies that contain the same or overlapping sets of participants can be identified
- Results for unpublished studies can be calculated and incorporated, thus reducing the problem of publication bias
- The statistical analysis can be standardised across studies (for example, the analysis method, how continuous variables are analysed, the time points assessed, and so on) and more appropriate or advanced methods can be applied where necessary
- Model assumptions in each study can be assessed, such as proportional hazards in Cox regression models, and complex relationships like time dependent effects can be modelled
- Estimates adjusted for baseline (prognostic) factors can be produced where previously only unadjusted estimates were available, which may increase statistical power and allow adjustment for potential confounding factors
- Baseline (prognostic) factors can be adjusted for consistently across studies
- Meta-analysis results for specific subgroups of participants can be obtained across studies (for example, those with a particular biomarker level), and differential disease stage effects can be assessed across individuals, which can help reduce between study heterogeneity
- Accounting for the clustering across studies can be considered, whether by using one-step or two-step meta-analysis.
- Accounting for the interaction between PF and the confounding factors can be considered.

8.2.3 How IMPF Projects can be Improved

Many aspects of the conduct and reporting of IMPF projects are done well. For example, all of the 20 articles I assessed provided a statistical methods section in their methods; all 20 reported the total number of participants within their IPD; 15 of the 20 accounted for heterogeneity in prognostic factor effects or justified why not; and 16 of the 20 considered the independent prognostic importance using a multivariable model. However, the available data from individual studies impose several challenges to derive a sensible summary estimate from an IMPF (see Chapter 2 Figure 2.8) and I also identified a number of ways IMPF projects can be improved (Figure 8.2). In particular, it was a surprise that protocols and ethics approval for IMPF projects were rarely mentioned. Protocols are essential components of any research project and enhance its credibility, so they should be made and referred to. In terms of ethics, researchers may have not considered this to be relevant if IPD was being used for the same objectives as in the original studies, for which ethical approval may exist.

Reporting standards in general must be improved within IMPF. Basic information was often missing, such the number of participants and events within each study providing IPD, and the keywords used to search for relevant studies. Researchers are encouraged to consider recent guidelines for reporting an IPD meta-analysis[60], which supplement existing reporting guidelines for meta-analysis of a non-IPD approach[70 172]. Of course, an improved reporting of primary studies according to the REMARK guidelines[23] is also needed, and this would be most helpful for several steps toward the IMPF project (e.g. examining study relevance according to inclusion/exclusion criteria; identifying outcomes considered and their specific definition; details about available variables and measurement techniques).

Figure 8.2: How can IMPF be improved

Rationale & Initiation

- Produce a protocol for the IMPF project prior to its initiation (detailing all aspects of rationale, conduct and statistical analysis) and reference this upon publication of the IMPF
- Consider whether ethics approval is necessary for the IMPF project, and report this upon publication

Process of obtaining IPD

- Report how primary study authors were approached to obtain their IPD
- Report the strategy used for searching the literature for relevant studies (if relevant), including keywords used and databases searched.
- Provide a flowchart showing the search strategy, classification of identified articles, and retrieval of IPD from relevant studies (where relevant)
- Consider how to improve retrieval of IPD from unpublished studies

Details of IPD obtained

- Report number of patients and events for each included study
- Report a summary of the missing data for each study
- Report the reasons why IPD was unavailable for some studies (if relevant), and report the number of patients, number of events and summary prognostic factor results in such studies

Type and quality of IPD obtained

- Consider and report the quality of studies for which IPD were obtained; in particular, are they all of comparable quality?

Statistical methods used

- Check and report the assumptions of the statistical models used; in particular, do model assumptions appear valid in each study separately?
- Where possible, analyse continuous factors on their continuous scale and consider non-linear trends
- In multivariable analyses, define the criteria used to decide whether a factor has independent prognostic value over other factors; also potentially consider whether the interaction between two (or more) prognostic factors is important
- In the meta-analysis, account for clustering of patients within studies (and do not merge IPD and analyse as if IPD all came from a single study) and report how this was done.
- Measure and, if necessary, account for between study heterogeneity in the prognostic factor effect(s) of interest when undertaking meta-analysis
- Where sufficient studies are available (e.g. 10 per covariate of interest) and heterogeneity exists, examine the potential causes of such heterogeneity.
- Consider whether meta-analysis conclusions change when restricting to IPD from the higher quality studies (if relevant)

Assessment of publication bias and availability bias

- Consider the potential impact of publication bias and availability bias on IPD meta-analysis results; in particular, are studies providing IPD comparable to those studies not providing IPD (if relevant)?

Reporting guidelines

- Utilize reporting guidelines for meta-analysis, such as those for MOOSE[70] and IPD meta-analysis[60]

In terms of statistical analysis, some IMPF projects chose to analyse continuous factors on a categorised scale without good reason. This approach has severe disadvantages as it loses statistical power and weakens the ability to assess non-linear prognostic factors effects[173]. When continuous factors are presented categorised within the available IPD itself, it is questionable whether a summary meta-analysis result is even sensible if different cut points are used across studies.[174] Another issue is that in five articles the meta-analysis method did not appear to account for clustering of participants within studies, and thus treated the IPD as if all coming from a single study, which is not appropriate. Further, four articles did not consider potential between-study heterogeneity in prognostic factor effects, even though heterogeneity is one of the most pertinent considerations in any meta-analysis. Also, missing data is considered as one of the substantial challenges. Among all of the 20 IMPF articles, each article considers at least one of the missing data problems such as missing data, missing variables and missing outcomes.

Another critical issue is properly acknowledging the heterogeneity of patient populations, treatment of the patients, and the available 'standard' prognostic variables used to adjust the effect of the factor of interest in each study. Issues of the heterogeneity concerning patient populations and treatment may be tackled by excluding subgroups of the patients from some of the single studies, by stratifying analyses or by restricting analyses to more homogeneous subgroups. Ideally the same adjustment variables should be used in each study, but one will be restricted by those available in the IPD provided and sometimes relevant factors may be missing, which should be clearly noted. For example, Thakkinstian et al[81]. note that they '*could adjust for only a few clinical factors. Other factors (e.g. immunosuppressive drugs and dosage, viral hepatitis infection, duration of dialysis, etc.) that were previously associated with poor outcomes after renal transplantation were not available in the datasets obtained.'*

Perhaps most importantly, publication bias and availability bias were rarely considered in the 20 IMPF articles examined. These problems are crucial to consider, as they may cause the IPD available to be a biased (non-random) portion of a potential prognostic factor's evidence base[55]. Publication bias is a well-known concern in meta-analysis, and having IPD from studies does not negate the fact that additional unknown, unpublished studies may also exist[157]. Funnel plots and tests for asymmetry ('small-study effects') are useful tools for identifying potential publication bias, but these were only used in two of the 20 articles. Availability bias is an added concern for IPD meta-analysis, and relates to when not all *known* studies provide their IPD. For example, IPD may be less obtainable from smaller studies and/or studies with non-significant findings as they are more likely to have lost or destroyed their IPD. In such situations, comparison of the summary results in IPD and non-IPD studies is useful to see if they are similar, as done by Rovers et al.[97] and Thakkinstian et al[81].

8.2.4 Should a Meta-analysis Account for the Clustering of Patients within-studies and is a One-step or Two- step more Appropriate?

One of the key challenges for IMPF articles was how meta-analysis should account for the clustering of patients within studies and whether one-step or two-step approach was more appropriate; in the reviewing for IMPF articles in chapter 2 there was only 14 out of 20 articles that accounted for the clustering of patients within studies. Further, it was clear that the statistical differences (if any) between one-step and two-step methods were not understood, as there is not a common used approach, and even accounting for clustering was not always considered important.

In chapter 4, I developed 11 IPD meta-analysis models for examining a single PF with a binary outcome; The difference in the model structure arise from using a one-step or two-step approach; accounting for the clustering of patients within studies by using two-step model or

one-step model[66 94 100 124]; whether one-step including indicator variable that reflect the impact of the clustering of patients across studies or ignoring the clustering of patients across studies by assuming the IPD studies as it came from one study; whether we account for the correlation between parameters in the second step of meta-analysis or not; also whether we account for the residual variation for the same values of the PF in each study (see Table 4-1).

The 4 IPD studies of the TBI dataset that investigate whether there is an association between age as a PF and probability of death by 6 months (see Table 1-2) was used to illustrate the methods. The results yielded *no* important difference across the 10 estimable models (recall that it was difficult to estimate the pooled effect size of the PF from model 1 as it is very difficult to estimate the residual variation, σ_i^2); in particular all of the 10 models have approximately the same pooled odds ratio and the same confidence interval, and all of the models yielded a P-value <0.05. This confirmed that there is a positive association between age and six month mortality rate (if age increase the probability of death increase) and age is still a PF. Initial conclusions suggested standard errors were smallest when correlation was accounted for, when residual variation was ignored, and when clustering was ignored. However smallest standard error does not necessary mean the best model. Indeed, the results did not help to identify the *best* meta-analysis model. Thus a simulation study was required to examine the effect of ignoring the clustering of patients across studies and whether one-step or two-step model more appropriate.

In chapter 5, I undertook a simulation study for three out of the eleven IPD meta-analyses models from chapter 4 to assess whether the clustering within studies can be ignored, and whether one-step or two-step approach is performed better. I considered these are the key questions to answer. In particular, I selected standard two-step model that account for the clustering of patients accross studies; one-step account for the clustering of patients across

studies by adding an indicator variable; and one-step ignoring clustering by assuming the IPD studies as it came from one study. I undertook my simulation for a single binary PF, and then a single continuous PF.

The simulation studies was undertaken for small and large number of IPD studies ; 5 and 10 studies repectively; also I assumed that the sample size in each IPD study could be small (30 to 100 patient) or large (30 to 1000) patients. The simulation study was repeated when the PF was continuous (by assuming normal distribution) and binary (by assuming Bernoulli distribution) . The outcome was binary (0 for alive and 1 for death) and it was generated for each patient according to a specified relationship between the outcome and the PF.

Bias, percentage bias, MSE, and coverage (95%) were recorded to assess the performance for the three IPD meta-analysis models. The findings were similar whether the PF was binary or continuous. The results of the simulation studies yielded that:

- There is *no* difference between the three IPD meta-analysis models when the Standard deviation for $\propto$ (the baseline risk) was 0 or 0.25.
- When the standard deviation for $\propto$ was large 1.5.
    - This produces inaccurate results for one-step model that ignore the clustering across studies (i.e. the coverage of the pooled estimate was considerably less than 95%, and it was downwardly biased).
    - An accurate result was seen for one-step model that account for the clustering of patients across studies (the coverage was around 95%, and there was no bias).
    - The results of two-step model yielded a too wide confidence interval with coverage above 95% when the PF effect was also large.

    To sum up, ignoring clustering across studies is not recommended[131 138 139], and the results of one-step accounting for clustering and two-step approach are often very

similar, only in a few situations with a binary PF did the two-step methods perform worst. This may be because the two-step model did not account for the correlation between parameters ($\propto_i$ and β_i) in the second step, and that the normality assumption in the second step fails. However, I would recommend one-step meta-analysis that account for the clustering across studies as it was consistently the best model across all simulation scenarios considered.

8.2.5 Benefit for Applying Fractional Polynomial Approach for Developing One-step and Two-step Models.

There are many issues for the analysis of continuous PFs; one of these issues arises when the assumption of linearity is found to be untenable, and I consider a fractional polynomial (FP) approach to address this problem[144 145 148]. Note that residual variation in models in chapter 4 highlight that a non-linear relationship in age is potentially more sensible for the TBI dataset. In chapter 6, the focus was on considering the possible two-step and one-step models for IPD studies that account for the clustering of patients across studies; in particular by using first and second FP order for IPD meta-analyses logistic regression models. The closed test was used to test the linearity assumption between the PF and the outcome, and it was also used to decide the best FP function (FP1 or FP2). TBI dataset was again used to estimate the pooled effect size of the PF for possible two-step and one-step models.

Two-step Approach

In two-step approach, I only fitted FP1 logistic regression model (p=1). However, it was difficult to decide the best fitted model that has the highest chi-square statistics (as there are 8 possible FP1 models), as there were different best powers for the 4 TBI trials when a 10 % significance level was used in the closed test. I handled this problem by adding up the 4 chi-square statistics for each model and I selected the model that has the highest sum chi-square

statistics (this occurred for age^2). Interestingly, however the closed test approach with 5% significance level gave age as a linear trend to be most appropriate.

One-step Approach

In one-step approach, two-models have been fitted; the first one was one-step FP1 IPD meta-analysis logistic regression model and it produced age^2 as the best fitted model. The second model was one-step FP2 IPD meta-analysis logistic regression model and it produced age^2 and age^{-2}, but the $\hat{\beta}_1$of age^{-2} was insignificant; thus the best model just includeage^2. The closed test approach gave age^2 as the best model regardless of the 5% or 10% significance level. Thus, at the 10 % level, it gives different conclusions to the two-step approach.

In conclusion, I would recommend one-step FP model over two-step FP model, as it is easier to apply and it forces the model to have the same power in each study, without need to all of the procedures that I considered in the two-step approach. Also closed test was a helpful tool to test the linearity assumption and help to choose between the simple and complex FP functions.

8.2.6 Assess the Impact of Small-study Effect on IMPF Articles.

Small-study effects may arise in IPD meta-analysis projects if only IPD from significant small-studies are available[104]. A contour funnel plot, cumulative funnel plot, trim and fill method and Egger's test were used to assess the small-study effect[106 158 166] in 6 IMPF articles; these articles were mainly selected according to the availability of their data (aggregated data or IPD studies). In particular, aggregated data were provided for 3 articles and the IPD were provided for the IMPACT project (relating to 3 articles specifically).

Generally, despite of array of many analyses, these was very little evidence for small-study effects; indeed in the IMPACT data, most small-study effects suggest that the PF effects will

be larger than those originally estimated (i.e. the opposite of a publication bias mechanism). In Look et al., however, there is a significant evidence for the small-study effects, but this was found unlikely to change the original conclusion for the PF (i.e. uPA), even though the pooled results was potentially smaller than thought.

The limitation here was that the assessment for small-study effects was only taken for 7 IMPF articles; this is mainly due to the difficulty to obtain suitable aggregated data or IPD for examining small-study effects. This is due to different reasons such as some studies provides the forest plot without writing the effect size and its confidence interval for each study, and I e-mailed the authors to ask them about the aggregated data or IPD data, but many of them did not respond of to my request. To sum up, there is very little evidence for the problem of publication bias or small-study effect in IMPF articles according to my limited review; even when small-study effects were a concern, the original conclusions were unlikely to change. This is encouraging.

8.3 Future Work

In Chapter 3 to 7 I examined some of the challenges listed in Figure 2.8. However, I have clearly not addressed all of them such as different methods of measurement and missing data. For example, in Trivella et al.[72] the authors mention that there is different methods of measurement for microvessel density; they overcome this problem by dividing the data into two groups (each group has the same method of measurement). In Look et al.[14] the authors also found different methods of measurement as they state that "*Levels of uPA and PAI-1 in tumour tissue extracts were determined by different immunoassays; values were ranked within each dataset and divided by the number of patients in that dataset to produce fractional ranks that could be compared directly across datasets"*. In the future work, I will consider the problem of missing data by using one of the statistical models such as

imputation analysis[175 176] and also I will try to examine the problem of different methods of measurement and generate some suggestions.

In Chapter 3, I compared between IPD and aggregated data meta-analysis approach, and I have shown by an empirical example that IPD meta-analysis is more favourable over aggregated data approach. However, I recognise that IPD is not always available. In the future work, I will examine the impact of combining aggregated data and IPD when some IPD are not available.

In Chapter 4, there are a number of limitations of the 11 IPD meta-analysis models such as the statistical analysis is applied to just a single PF (univariate analysis) and I assumed fixed-effect method across studies as that there was no heterogeneity across the TBI ($I^2 = 0\%$). In the future work, I will consider model extension for accounting for the heterogeneity across studies. Also, I will consider the multivariable analysis by adjusting the PF of interest by other Prognostic factors[177].

In Chapter 5, there are some limitations for the simulation studies. First, I applied the simulation on a single variable only (continuous or binary PF); and I did not consider the multivariable model. Second, I assumed the linearity assumption for the continuous variable. Finally, I applied the simulation study only on three possible models for IPD meta-analysis and I assumed fixed effect across studies. In my future research, I will consider the following points: (i) the multivariable IPD meta-analysis models; (ii) random-effects to account for heterogeneity across studies (iii) and methods to account for the correlation between parameters by using multivariable random-effects two-steps meta-analysis models.

In Chapter 6, generally, two-step FP IPD meta-analysis models, in the first step, allows a different FP model in each study but this makes it difficult to synthesize in the second step.

On the other hand, one-step model forces each study to have the same FP model (through estimates could be different, they all have same shape, e.g. all quadratic or all cubic etc.). Saurberi et al., (personal communication) suggest a new approach for FP two-step models, which allow a different FP model in each study and then averages these across studies using a two-step approach. Also random- effects approach it would be interesting to compare their approach to my one-step approach in the future work.

In Chapter 7, I assessed the small-study effects for 6 IPD studies, and I found that there is very little concern for the small-study effects in IMPF. In the future work, more IMPF articles need to be assessed, especially in situations where IPD are unavailable for some studies. Then, I need to examine aggregated data and IPD. More simulations looking at the performance of Eggers D-var method for PF meta-analysis is also needed.

8.4 Conclusion

Although IPD offer many advantages over aggregated data. It still presents many challenges for meta-analysis of PFs. In this thesis, I have tackled some of these issues, but much research is needed in the coming years to help ensure IPD meta-analysis produces reliable evidence-based PF results. My key conclusion are shown in Figure 8.3

In this thesis, I only used and studied retrospective IPD meta-analysis of PFs, due to the ease and quickness of using existing data. However they are restricted by the data already collected by primary study (e.g. certain variables and outcomes may be missing)[1]. Prospective IPD meta-analysis of PFs may reduce many of the pitfalls that I found for the retrospective IPD meta-analysis, as in prospective studies, patients are recruited at the current date and followed for an adequate length of time to identify what outcomes are achieved in all studies simultaneously. This allows the same PFs, baseline characteristics and outcomes to be recorded by the researcher, and to ensure similar methods of

measurement, data coding, and variable inclusion from study to study. Thus, ultimately the gold standard IPD approach may actually be a prospective meta-analysis.

Figure 8. 3: Key Conclusions from this thesis

- IPD meta-analysis hugely beneficial for meta-analysis of PF studies especially compared to aggregated data approach.
- Yet IPD still presents many methodological, statistical and practical challenges.
- Often there is very little difference between one-step, two-step and even models that ignore clustering.
- But when there is large variation in baseline risk, one-step IPD meta-analysis models that account for clustering are preferred; ignoring clustering leads to a low coverage, and the two-step (occasionally) leads to too high a coverage. Thus one-step accounting for clustering in my recommended best model.
- When modeling a continuous PF and assuming a linear trend, examining the residual variation due to this factor helps decide whether as non-linear trend is sensible.
- FP modeling allows non-linear PF trends; again the one-step framework is preferred because it forces each study to have the same FP terms.
- Small-study effects may arise for IMPF but there was little evidence of this when examined. Only in one dataset (Look et al.) was it a major concern, but even then clinical conclusion is unlikely to alter after adjusting for funnel plot asymmetry.

APPENDIX

Appendix A

Appendix A1: Review of IPD meta-analyses of prognostic factor studies

Aim: To critically examine published IPD meta-analyses of PF studies (in relation to any outcome) and ascertain:

- their objectives and rationale
- how and why they were conducted
- the process used to obtain IPD and the success/problems therein
- how the quality of primary studies was assessed
- what statistical methods were used and the problems therein; in particular how heterogeneity was examined, how many adjustment factors were used and how they were modelled, and how continuous factors were assessed
- how publication bias related issues were examined
- what the limitations and successes of the IPD project were, and the benefits/differences over aggregate data meta-analysis

- *We call a PF study one that looks to see if one or more factor is associated with an outcome in patients with existing disease at baseline.*
- We are NOT looking at *predictive markers* (i.e. those predicting response to treatment) and NOT *prognostic models* (i.e. those developing a risk score to predict outcome in future individuals). But of course the papers may have assessed the PFs alongside or enroute to a prognostic model or treatment predictions
- In our review, all the questions relate to the prognostic factor assessments. Other parts of the paper (e.g. treatment effect, prognostic models etc) are not of interest at all.

Below is a list of questions we will ask of a sample of the IPD meta-analyses identified

A.1.1 General questions

- What country is the corresponding author located in? (i.e. what is the central location for the IPD project?)
- is there reference to a protocol for the IPD project, and, if so, were details given as to where it can be found?
- Was ethics approval granted for the IPD project? If not, were reasons given as to why ethics approval was not necessary and, if so, what were the reasons?
- What were the key prognostic factor research aims of the paper? In particular, are they looking to see if a prognostic factor is important on its own, or whether it is important after adjusting for existing PFs?
- What was the main motivation to start the IPD project? (e.g. is it to resolve current disagreements; or to look at a new novel marker; is it connected to

financial support from a funder or a company or an HTA assessment?)
- At baseline what was the condition of the patients being assessed (e.g. what disease
- did they have, or what operation had they just had etc)
- What PFs were of key interest?
- What outcomes were of interest?

A.1.2 Process of obtaining IPD

- What were the stated reasons given for taking the IPD approach, rather than a meta-analysis of aggregate data?
- Was another meta-analysis approach tried before and, if so, why was this IPD project needed in addition to this previous meta-analysis
- If a meta-analysis of aggregate data had been done before, what were the stated conclusions and limitations of the research?
- What was the process used to identify relevant studies for the IPD meta-analysis? (e.g. literature review, or collaborative group).
- If literatures review, then what search strategy was used (e.g. search of Medline, Embase using keywords)?
- If collaborative group, how were studies chosen to be included in the collaborative group (e.g. uPA is based on projects funded from an EU grant, so all EU studies included)?
- What were the inclusion / exclusion criteria for deciding from which studies IPD would be sought?
- For those studies for which IPD was desired, were there any requirements on the information needed within the IPD? In particular, were there any requirements on studies providing a certain set of 'adjustment factors'? (e.g. did they say that tumour size or stage of disease had to be available in the IPD for the study to be worth providing the IPD?)
- How were authors of relevant studies approached for IPD (e.g. e-mail, letter, phone etc.)? How many authors replied on first or second request?
- Are there any details on the cost and time required to obtain and 'clean' the IPD provided?
- How many studies (or collaborating groups) were ultimately approached for IPD, and what proportion of these studies/groups actually provided IPD?
- (If appropriate) what were the reasons given as to why some studies refused to provide IPD?

A.1.3 Details of the IPD obtained

- Did those studies that provided IPD give all their IPD or only a proportion?
    - If only a proportion, then what was omitted (e.g. certain patients, certain variables etc) and why?
- Were details given of any differences in the IPD provided to that used in one or more previous publications? If yes, what were they (e.g. IPD has a longer follow-up time etc)
- How many studies ultimately provided IPD?
- Was the number of patients within each of the IPD studies given

- Was the number of events given for each outcome within each of the IPD studies?
- Were details of any missing individual-level data within the available IPD given for each study, and, if so, what were they?
- How many variables were considered as adjustment factors in the original studies (i.e. aside from the prognostic factor(s) of interest, how many other patient-level variables were available in each study)?
- Did all studies have all variables, or was there missing variable information across studies?

A.1.4 Quality of included primary studies

- Did the inclusion / exclusion criteria for an IPD study include an assessment of study quality?
    - If yes, what quality criteria were used to decide inclusion or exclusion (or 'low' quality and 'high' quality)?
    - Were IPD still sought from low quality studies?
- If the inclusion / exclusion criteria did not involve an assessment of study quality, was a quality assessment done at some other stage (e.g. for those studies actually providing IPD)?
    - If yes, describe when the assessment took a 'low quality study' and a 'high quality study'
- How were 'low quality studies' dealt with? In particular, was IPD sought from these studies and, if so, how were they included in the meta-analysis (if at all)?

A.1.5 Statistical methods for an IPD meta-analysis

- Was a statistical analysis plan given or mentioned in the Methods section?

Meta-analysis approach

- Were the data meta-analysed by:

  (a) Lumping all the data together into one big dataset, and ignoring clustering by trial or collaborative group; or
  (b) A one-step analysis, where the data from all studies/collaborative groups are analysed together but with clustering by study/group accounted for; or
  (c) A two-step approach, where the data are first analysed separately in each study, and then their model estimates are pooled together in second-step.

- What type of statistical models was used? In the two-step approach, details are needed here of how individual studies were analysed, and then how the model estimates were pooled using meta-analysis. In the one-step approach, again details are need here of the one-step model itself (Cox regression, logistic regression) and the meta-analysis assumptions therein (e.g. fixed or random-effects on the prognostic factor effect)

Heterogeneity

- Was the presence of between-study heterogeneity assessed and, if so, how? (e.g. I2 statistic)

- Did the meta-analyses account for between-study heterogeneity? If so, how? (e.g. use place, and the quality criteria used to define of random-effects)
- Were the factors causing between-study heterogeneity investigated? If so, how? (e.g. including study-level covariates in the model, or performing meta-regression)

Continuous factors

- How were continuous PFs analysed, on a continuous scale or categorized? If categorised
- Were reasons given as to why this was done?
- How many cut-points were used, and how were they chosen? If on a continuous scale
- Were non-linear trends assessed and, if so, how were they modelled? (e.g. spines, fractional polynomials)
- Was the continuous factor analysed on its original scale, or was it on a transformed scale?
- If a one-step meta-analysis was used, was the continuous factor cantered about the mean value in each study?

Clinical factors

- Were clinical factors adjusted for in the model; if so, what were they?
- For the prognostic factor(s) of primary interest were different methods of measurement used across studies. If so, how did the authors address in their analyses (if at all)?
- Did each patient receive the same treatment? If not, how was treatment received ac- counted for in the analyses (if at all)? (NB again potential distinction needed here of how individual study analyses accounted for treatment, and then at the meta-analysis level how treatment was accounted for) - essentially look for any details of how and when treatment was incorporated into the analyses that are described.

A.1.6 Analyses of multiple PFs together and adjustment factors

- Was the prognostic factor assessment adjusted for other variables (e.g. other PFs or adjustment variables)? If so:
- What method was used (e.g. Cox regression)? - Were combinations of PFs assessed (i.e. was the interaction between two or more PFs modelled). If so, how?

- If multivariable models were fitted (i.e. models that included multiple variables)

  - What criteria were used to decide inclusion in the model? (e.g. statistical criteria, such as $p < 0.1$, or clinical criteria such as a hazard ratio > 2 or inclusion of 'smoking' variable regardless).
  - Were results for non-significant presented in full (i.e. an effect estimate and its confidence interval, or p-value). - How were continuous factors analysed in the model.

A.1.7 Missing data and check of assumptions

- how were missing data handled at the study-level (e.g. missing treatment information) and/or at the patient-level (e.g. missing prognostic factor values)
- Were the assumptions of the statistical models validated (e.g? proportional hazards) within each study
- What problems limited the statistical analysis (e.g. different method of measurements, different set of adjustment factors available etc) and how did the authors attempt to overcome these problems?

A.1.8 Bias in the set of IPD available / differences between IPD and non-IPD studies

- Was there an assessment or discussion of whether the IPD studies were a biased set of all studies (akin to the problem of publication bias), e.g. using a funnel plot? If yes, describe the assessment and what was found.
- (if appropriate) for studies not providing IPD, were details given as to the number of patients and events in these studies
- (If appropriate) were there any other details provided on the qualitative or quantitative differences between those studies providing IPD and those studies not providing IPD? If so, what were these differences?
- (If appropriate) were details given regarding the robustness of meta-analysis results to the inclusion/exclusion of non-IPD studies? If so, what were these differences?
- (If appropriate) were IPD and non-IPD studies meta-analysed together? If so, what methods were used to do so?

A.1.9 Reporting and Discussion

- In terms of reporting, was it stated anywhere that the reporting guidelines of QUO-RUM or MOOSE were used, or was reference given to the QUOROM or MOOSE articles? (see refs below)
- What were the main conclusions of the IPD project in the Discussion?
- How do the conclusions differ (if at all) to any previous meta-analyses that have been done on the topic?
- What strengths of the IPD project were noted in the Discussion?
- What limitations and problems of the IPD project were noted in the Discussion?
- What is the key further research?

Appendix B

Appendix B1: The proof of $cov(\hat{\beta}_0, \hat{\beta}_1) = -\text{var}(\hat{\beta}_0)$, by using logistic regression model is explained as follows:

Note that $\hat{\beta}_0 = \hat{\alpha}$, and $\hat{\beta} = \hat{\beta}_1$. Let Y be a dichotomous random variable denoting the outcome of the PFs, and let $X = x_1, x_2, ..., x_{p-1}$ be a collection of PFs (for simplicity, I assume that there is only one PF); the conditional probability of the outcome is presented by $P(Y = 1/x) = \pi(x)$, where $\pi(x)$ has the form

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}$$

If the values of the PF are varied from participants (i=1 to n), then $\pi(x)$ can be written as:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1})}{1 + \exp(\beta_0 + \beta_1 x_{i1})}$$

The logistic regression problem is then to obtain an estimate of the vector:
$\hat{\beta} = (\beta_0, \beta_1)$
As with logistic regression, the matrix X for one PF and the intercept can be written as:

$$\begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ ... & ... \\ 1 & x_{n1} \end{bmatrix}$$

This matrix is called the regression matrix, where 1's denoted to the intercept, and x_{ij} refer to the PF.
Maximum likelihood method is used to estimate the vector $\hat{\beta}$, and the likelihood function given as:

$$l(\beta) = \prod_{i=1}^{n} \pi_{i=1}^{y_i} (1 - \pi_i)^{1-y_i}$$

, and the log likelihood function is given as:

$$L(\beta) = \sum_{i=1}^{n} [y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)]$$

Estimates for the variances and covariances of the estimated parameters β_j are computed using the following equations. Let $\hat{I}(\hat{\beta}) = X'VX$, where X is the $n \times p$
Matrix called regression matrix, and V is a $n \times n$ diagonal matrix with i^{th} diagonal term $\pi_i(1 - \pi_i)$. That is, the matrix X is:

$$\begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \dots & \dots \\ 1 & x_{n1} \end{bmatrix}$$

The matrix X' is

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \end{bmatrix},$$

And the matrix V is:

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix},$$

Now I will multiply the first two matrixes $X'V$ and the result will be

$$X'V = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} \dots & & x_{n1} \end{bmatrix} \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & \hat{\pi}_2(1-\hat{\pi}_2) & \dots\dots\hat{\pi}_n(1-\hat{\pi}_n) \\ x_{11}\hat{\pi}_1(1-\hat{\pi}_1) & x_{21}\hat{\pi}_2(1-\hat{\pi}_2) \dots\dots & x_{n1}\hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix},$$

Then $X'VX$ is

$$X'VX = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & \hat{\pi}_2(1-\hat{\pi}_2) & \dots\dots\hat{\pi}_n(1-\hat{\pi}_n) \\ x_{11}\hat{\pi}_1(1-\hat{\pi}_1) & x_{21}\hat{\pi}_2(1-\hat{\pi}_2) \dots\dots & x_{n1}\hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix} \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \dots & \dots \\ 1 & x_{n1} \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1)+\hat{\pi}_2(1-\hat{\pi}_2)+\dots+\hat{\pi}_n(1-\hat{\pi}_n) & x_{11}\hat{\pi}_1(1-\hat{\pi}_1)+x_{21}\hat{\pi}_2(1-\hat{\pi}_2)+\dots+x_{n1}\pi_n(1-\hat{\pi}_n) \\ x_{11}\hat{\pi}_1(1-\hat{\pi}_1)+x_{21}\hat{\pi}_2(1-\hat{\pi}_2)+\dots+x_{n1}\pi_n(1-\hat{\pi}_n) & x_{11}^2\hat{\pi}_1(1-\hat{\pi}_1)+x2_{21}\hat{\pi}_2(1-\hat{\pi}_2)+\dots+x_{n1}^2\pi_n(1-\hat{\pi}_n) \end{bmatrix}$$

The inverse of the $X'VX$ matrix can be calculated as follows:
First, the determine of $X'VX$ denote by $\det(X'VX)$, and is given as

$$\det(X'VX) = [\hat{\pi}_1(1-\hat{\pi}_1)+\hat{\pi}_2(1-\hat{\pi}_2)+\dots+\hat{\pi}_n(1-\hat{\pi}_n)]\,[\,x_{11}^2\hat{\pi}_1(1-\hat{\pi}_1)+x2_{21}\hat{\pi}_2(1-\hat{\pi}_2)+\dots+x_{n1}^2\pi_n(1-\hat{\pi}_n)]$$
$$-[x_{11}^2\hat{\pi}_1(1-\hat{\pi}_1)+x2_{21}\hat{\pi}_2(1-\hat{\pi}_2)+\dots+x_{n1}^2\pi_n(1-\hat{\pi}_n)]^2$$

Second, the inverse of $X'VX$ is called the variance covariance matrix of the estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_1$, this is given as:

$$\frac{1}{\det(X'VX)}\begin{bmatrix} x_{11}^2\hat{\pi}_1(1-\hat{\pi}_1)+x_{21}^2\hat{\pi}_2(1-\hat{\pi}_2)+....+x_{n1}^2\pi_n(1-\hat{\pi}_n) & -[x_1\hat{\pi}_1(1-\hat{\pi}_1)+x_2\hat{\pi}_2(1-\hat{\pi}_2)+....+x_n\hat{\pi}_n(1-\hat{\pi}_n)] \\ -[x_1\hat{\pi}_1(1-\hat{\pi}_1)+x_2\hat{\pi}_2(1-\hat{\pi}_2)+....+x_n\hat{\pi}_n(1-\hat{\pi}_n)] & \hat{\pi}_1(1-\hat{\pi}_1)+\hat{\pi}_2(1-\hat{\pi}_2)+....+\hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

Appendix B2: Special case: Bivariate Random Meta-Analysis (BRMA) revert to Univariate Fixed effect Meta-Analysis (UFMA) for $\hat{\alpha}$ when both of outcomes and PF are binary data, the correlation within study does not effect on the estimated value of $\hat{\alpha}$

Note that, let $var\left(\hat{\beta}_{i1}\right) = S_{i1}^2$ and $\left(\hat{\beta}_{i2}\right) = S_{i2}^2$,$\lambda_i = cov\left(\hat{\alpha}_i, \hat{\beta}_i\right)$, $\hat{\propto}_i = Y_{i1}$, and $\hat{\beta}_i = Y_{i2}$
BRMA is a statistical model that can be utilised to estimate $\hat{\beta}_j$; it also includes the effect of the within and between studies correlation ; again assume that each study summary statistic Y_{ij} (where i=1 to n and j=1 to 2 outcomes) is assumed an estimate of a different underlying true value θ_{ij} in each study; each θ_{ij} is assumed to be drawn from a distribution with mean value β_j and between study variance $\hat{\tau}_j$; then the BRMA model can be written mathematically as follows:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix}, \boldsymbol{\delta}_i\right) \qquad \boldsymbol{\delta}_i = \begin{pmatrix} s_{i1}^2 & \lambda_i \\ \lambda_i & s_{i2}^2 \end{pmatrix} \tag{5}$$

$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{\Omega}\right) \qquad \boldsymbol{\Omega} = \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix}$$

Where δ_i and $\boldsymbol{\Omega}$ are the within-study and between-study covariance matrices respectively; λ_i is the within-study covariance, τ_{12} is the between-study covariance. As can be seen in equation (6), the BRMA consider the correlation within study and between studies. The covariance within-studies assume to be known. On the other hand, the UFMA does not account for the within and between studies correlation.
BRMA model revert to UFMA model when $\hat{\tau}_1^2 = \hat{\tau}_2^2 = \hat{\tau}_{12}^2 = 0$ (i.e. there is no heterogeneity across studies and no correlation across studies), and $\lambda_i = -\, s_{i1}^2$
(i.e. the covariance within studies equal negative the variance of the $\hat{\alpha}$ when the outcome and the PF are binary data by using BRMA of logistic regression model). Note that the prove of $\lambda_i = -s_{i1}^2$ is explained in appendix B1. Also, there is only one case when BRMA revert to UFMA model for $\hat{\alpha}$ only, when $\hat{\tau}_1^2 = \hat{\tau}_2^2 = \hat{\tau}_{12}^2 = 0$, $\lambda_i = -s_{i1}^2$ and when the outcome is binary and the PF is binary. However, if the PF is continuous then the estimated value for $\hat{\alpha}$ by using UFMA model does not equal the estimated value for $\hat{\alpha}$ when the PF is continuous. Note that the focus here is on the estimated value for $\hat{\beta}$ (not on $\hat{\alpha}$).
First, the pooled estimate of the PF effect size of the BRMA can be estimated by using restricted iterative generalized least square method (RIGLS) as:

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} \left[\frac{Y_{i1}}{(\hat{\tau}_1^2 + s_{i1}^2)(\hat{\tau}_2^2 + s_{i2}^2) - (\hat{\tau}_{12} + \lambda_i)^2} \left[\sum_{k=1}^{n} \frac{(\hat{\tau}_2^2 + s_{i2}^2)(\hat{\tau}_1^2 + s_{k1}^2) - (\hat{\tau}_{12} + \lambda_i)(\hat{\tau}_{12} + \lambda_k)}{(\hat{\tau}_1^2 + s_{k1}^2)(\hat{\tau}_2^2 + s_{k2}^2) - (\hat{\tau}_{12} + \lambda_k)^2} \right] \right] + \sum_{i=1}^{n} \left[\frac{Y_{i2}}{(\hat{\tau}_1^2 + s_{i1}^2)(\hat{\tau}_2^2 + s_{i2}^2) - (\hat{\tau}_{12} + \lambda_i)^2} \left[\sum_{k=1}^{n} \frac{(\hat{\tau}_{12})(s_{i1}^2 - s_{k1}^2) + \lambda_k(\hat{\tau}_1^2 + s_{i1}^2) - \lambda_i(\hat{\tau}_1^2 + s_{k1}^2))}{(\hat{\tau}_1^2 + s_{k1}^2)(\hat{\tau}_2^2 + s_{k2}^2) - (\hat{\tau}_{12} + \lambda_k)^2} \right] \right] \right)}{\sum_{i=1}^{n} \frac{\hat{\tau}_1^2 + s_{i1}^2}{(\hat{\tau}_1^2 + s_{i1}^2)(\hat{\tau}_2^2 + s_{i2}^2) - (\hat{\tau}_{12} + \lambda_i)^2} \sum_{i=1}^{n} \frac{\hat{\tau}_2^2 + s_{i2}^2}{(\hat{\tau}_1^2 + s_{i1}^2)(\hat{\tau}_2^2 + s_{i2}^2) - (\hat{\tau}_{12} + \lambda_i)^2} - \left(\sum_{i=1}^{n} \frac{(\hat{\tau}_{12} + \lambda_i)}{(\hat{\tau}_1^2 + s_{i1}^2)(\hat{\tau}_2^2 + s_{i2}^2) - (\hat{\tau}_{12} + \lambda_i)^2} \right)^2} \quad (6)$$

, where $k = 1,....., n$ representing the n studies, and k is used to distinguish the summation from 1 to n within the summation for $i = 1,....., n$.

Second, by substituting in equation (5) when $\hat{\tau}_1^2 = \hat{\tau}_2^2 = \hat{\tau}_{12}^2 = 0, \quad \lambda_i = -s_{i1}^2$, then $\hat{\beta}_1$ can be written as:

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} \left[\frac{Y_{i1}}{(s_{i1}^2)(s_{i2}^2) - (-s_{i1}{}^2)} \left[\sum_{k=1}^{n} \frac{(s_{i2}^2)(s_{k1}^2) - (-s_{i1}{}^2)(s^2{}_{k1})}{(s_{k1}^2)(s_{k2}^2) - (-s_{k1}^2)^2} \right] \right] + \sum_{i=1}^{n} \left[\frac{Y_{i2}}{(s_{i1}^2)(s_{i2}^2) - (-s_{i1i}^2)^2} \left[\sum_{k=1}^{n} \frac{-s_{i1}^2(s_{i1}^2) - s_{i1}^2(s_{k1}^2))}{(s_{k1}^2)(s_{k2}^2) - (-s_{i1}^2)^2} \right] \right] \right)}{\sum_{i=1}^{n} \frac{s_{i1}^2}{(s_{i1}^2)(s_{i2}^2) - (-s_{i1}^2)^2} \sum_{i=1}^{n} \frac{s_{i2}^2}{(s_{i1}^2)(s_{i2}^2) - (-s_{i1}^2)^2} - \left(\sum_{i=1}^{n} \frac{(-s_{i1}^2)}{(s_{i1}^2)(s_{i2}^2) - (-s_{i1}^2)^2} \right)^2} \quad (7)$$

Third, by cancelling s_{i1}^2 out from the denominator (in the very down line), equation 6 reduce to:

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} \left[\frac{Y_{i1}}{(s_{i1}^2)(s_{i2}^2) - (-s_{i1}^2)^2} \left[\sum_{k=1}^{n} \frac{(s_{i2}^2)(s_{k1}^2) - (-s_{i1}{}^2)(s^2{}_{k1})}{(s_{k1}^2)(s_{k2}^2) - (-s_{k1}^2)^2} \right] \right] \right)}{\sum_{i=1}^{n} \frac{1}{(s_{i2}^2) - (-s_{i1}^2)} \sum_{i=1}^{n} \frac{s_{i2}^2}{(s_{i1}^2)(s_{i2}^2 - s_{i1}^2)} - \left(\sum_{i=1}^{n} \frac{1}{(s_{i2}^2) - (-s_{i1}^2)^2} \right)^2} \quad (8)$$

Fourth, by taking s_{k1}^2 as a factor from the numerator and $\sum_{i=1}^{n} \frac{1}{s_{i2}^2 - s_{i1}^2}$ as a factor from the denominator, equation 7 can be written as:

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} \left[\frac{Y_{i1}}{s_{i1}^2(s_{i2}^2 - s_{i1}^2)} \left[\sum_{k=1}^{n} \frac{s_{k1}^2(s_{i2}^2 - s_{i1}^2)}{s_{k1}^2(s_{k2}^2 - s_{k1}^2)} \right] \right] \right)}{\sum_{i=1}^{n} \frac{1}{s_{i2}^2 - s_{i1}^2} \left[\sum_{i=1}^{n} \frac{s_{i2}^2}{s_{i1}^2(s_{i2}^2 - s_{i1}^2)} - \sum_{i=1}^{n} \frac{1}{s_{i2}^2 - s_{i1}^2} \right]} \quad (9)$$

Fifth, by cancelling s_{k1}^2 and $s_{i2}^2 - s_{i1}^2$ out from the top numerator, and cancel $\sum_{i=1}^{n} \frac{1}{s_{k2}^2 - s_{k1}^2}$ with $\sum_{i=1}^{n} \frac{1}{s_{i2}^2 - s_{i1}^2}$ out, then equation 8 is given as:

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} \left[\frac{Y_{i1}}{s_{i1}^2} \right] \right)}{\left[\sum_{i=1}^{n} \frac{s_{i2}^2}{s_{i1}^2 \left(s_{i2}^2 - s_{i1}^2 \right)} - \sum_{i=1}^{n} \frac{1}{s_{i2}^2 - s_{i1}^2} \right]} \quad (10)$$

This can be written as:

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} \left[\frac{Y_{i1}}{s_{i1}^2} \right] \right)}{\left[\sum_{i=1}^{n} \frac{\frac{s_{i2}^2}{s_{i1}^2} - 1}{\left(s_{i2}^2 - s_{i1}^2 \right)} \right]} = \frac{\sum_{i=1}^{n} \frac{Y_{i1}}{s_{i1}^2}}{\sum_{i=1}^{n} \frac{1}{s_{i1}^2}} \quad (11)$$

As we can see from the equation (9) that the BRMA revert to UFMA
Note that, this prove only for the case that the outcome and the PF are binary; an applied example will be given later. However, this can be extended to the continuous case; in particular when the PF is continuous (e.g. age), but I will show that by applying on the TBI data set later.

Appendix C

Appendix C1: STATA codes for possible two-step IPD meta-analysis models

\\ **Model 2** accounts for the clustering across studies and did not account for within-study variation for the patients who have the same age; the pooled estimate of the PF is estimated through two steps\\

The first code:

by studyid: logit dead age10

The second code:

metan beta s.e. of beta, fixedi

Model 3: This model account for the clustering across studies and the variation within-studies for the patients who have the same age value in each IPD study. The pooled estimate effect size of the PF is estimated through two steps\\

The first code:

by studyid: xtmelogit dead age10, || age10: cov(indep) intp (1) var

The second code:

metan beta s.e of beta, fixedi

the explanation here is the same as in the model 2 except that this code ' || **age10: , cov(indep) intp(1) var'** is allowed the PF (age) to vary across patients with different value of age and **cov(indep)** means one variance parameter per random effect. The second step estimates the pooled effect size of meta-analysis by using 'metan'. The interpretation here is similar to model 2.

Appendix C2: STATA codes for possible one-step IPD meta-analysis models

\\ Model 6 ignores the clustering across studies and within study variation (assume that all of the data came from one study)\\

logit dead age10

\\ Model7 ignores the clustering across studies by including the indicator variables and accounting for the variation within patients who have the same age value\\

Xtmelogit dead age10 || age10:, cov (indep) intp (1) var

\\ Model 8 accounts for the clustering across studies and ignoring the variation within-studies for the patients who have the same value of the PF in each study\\

logit dead f11 f12 f13 f14 age10, nocons

\\Model 9 accounts for the clustering across studies by including indicator variables and account for the variation within-study for the patients who have the same value of the PF\\

xtmelogit dead age10 f11 f12 f13 f14, nocons || age10:, cov (indep) intp(1) var

\\ Model 11 accounts for the clustering across studies and ignoring the variation within-studies for the patients who have the same value of age in each study and assign a normal distribution for α to allow for different values for alpha in each IPD study\\

xtmelogit dead age10 || studyid: , cov(indep) intp(1) var

\\Model 11 accounts for the clustering across studies and account for the variation within-studies for the patients who have the same value of age in each study and assign a normal distribution for α to allow for different values for alpha in each IPD study\\

xtmelogit dead age10 || studyid: , || age10: f11 f12 f13 f14, nocons cov(indep) intp(1) var

Appendix C3: SAS code for bivariate fixed-effect (Model 4 and 5)

SAS code to fit a bivariate fixed-effect model in the second step and allowing a different alpha in each study, for the continuous age and after accounting for residual variation in the first step

```
data tbi;
input id    estimate    alpha beta  study1      study2      study3
      study4      var;
cards;
1     -2.199809   1     0     1     0     0     0     0.057277796
1     0.3105692   0     1     1     0     0     0     0.004113465
2     -1.928625   1     0     0     1     0     0     0.043479673
2     0.2672432   0     1     0     1     1     0     0.002791823
3     -1.956439   1     0     0     0     1     0     0.067149704
3     0.376585    0     1     0     0     0     0     0.003082048
4     -2.323136   1     0     0     0     0     1     0.122192753
4     0.3283813   0     1     0     0     0     1     0.007850244
;
run;
/* Bivariate meta-analysis allowing alpha to be different in each study
(and not estimated) but allowing beta to be the same in each study and
estimated */
proc mixed cl method=reml data=tbi ;
class id ;
model estimate = alpha beta / noint s cl ddf=10000 10000;
random alpha beta / subject=id g type=arh(1);
repeated / type=un subject=id group=id ;
parms
0.1
0
0
0.057277796
-0.01440608
0.004113465
0.043479673
-0.01018656
0.002791823
0.067149704
-0.01320622
0.003082048
0.122192753
-0.02883057
0.007850244
/ eqcons = 2 to 15;

run;
```

Appendix D

Appendix D1: The simulation STATA code program for binary PF.

```
** parameter setting from here
clear
clear matrix
set memory 1000m
set more off
set seed 12345678
cd "C:\my thesis" tempname simparm   /*record all parameter and population value for each senario*/
postfile `simparm' senario inercept_m inercept_v slope lymphnod n_studies using C:\my thesis\senario_monitor_cat.dta, replace
*****set different parameter here, following are Ghada's code
local a_m "-1.27"      /*intercept mean for control,you can set value list here, e.g. -1 0 +1*/
local a_v "0.25 "      /*intercept SD for control,you can set value list here, e.g. -1 0 +1*/
local b "0.1 0"   /*treatment effect, you can set value list here, e.g. -1 0 +1*/
local p_lym "0.2" /* prevenlance lymphnode positive, you can set value list here, e.g. 0.1 0.2 0.3*/
local n_studies "5 10" /* number of IPD_j in each meta analysis, you can set value list here, e.g. 5 10 20 30 */global repl = 1000   /*decide how many replication used*/
****** DO not change following code******
local senario = 1
foreach intpt_m in `a_m' {
        foreach intpt_v in `a_v' {
                foreach slp in `b' {
                        foreach lymph in `p_lym' {
                                foreach n_meta in `n_studies' {

                                quietly replic `intpt_m' `intpt_v' `slp' `lymph' `n_meta'   `senario'   /*simulate data*/

                                post `simparm' (`senario') (`intpt_m') (`intpt_v') (`slp') (`lymph') (`n_meta') /*save parameter setting for each senario*/

                                *Analysis data here or seperately?
                                use "C:\my thesis\data\senario_`senario'", clear
                                two_step `intpt_m' `intpt_v' `slp' `lymph' `n_meta'   `senario' /*runing two_step analysis*/

                                use "C:\my thesis\data\senario_`senario'", clear
                                step_cluster `intpt_m' `intpt_v' `slp' `lymph' `n_meta'  `senario' /*runing one_step cluster*/

                                use "C:\my thesis\data\senario_`senario'", clear
                                step_dummy  `intpt_m'  `intpt_v'  `slp'  `lymph'  `n_meta'    `senario'  /*runing one_step dummy*/

                                 conbin_result `intpt_m' `intpt_v' `slp' `lymph' `n_meta'   `senario' /*combine 3 kind of result together for same senario*/

                                local ++senario
                                        }/*end of study number loop*/

                                }/*end of lymphnode prevelence loop*/
                        }/*end of slope loop*/
                }/*end of  alpha variance loop*/
        } /*end of  intercept loop*/
        postclose `simparm'  /*close to save senario parameter*/

        ************
*JAN-21-2011
* programe here are called by file My thesis_leices
* programm are
```

```
*IPD_j
*replic
*two_step
*step_cluster
*step_dummy
*binary covariate
**********************
capture program drop IPD_j
capture program drop replic
capture program drop two_step
capture program drop step_cluster
capture program drop step_dummy
capture program drop conbin_result
program define IPD_j
        local alpha_m `1'
        local alpha_v `2'
        local beta `3'
        local p_x `4'

        clear /*clear the file in memory for new simulation*/
        local alpha= (rnormal(`alpha_m', `alpha_v')) /* alpha is normal dist value*/

        local n_ipd_j = int(30+(runiform()*(100-30))) /*30~100, Gadha subjects for each primary study_j*/
        quietly set obs `n_ipd_j' /*how many people in each IPD study*/

        gen x = rbinomial(1, `p_x')  /*x is bianry treatment variable, p_x is the percentage of X=1*/

        gen y = runiform() < invlogit(`alpha'+ `beta'*x) /*Thompson code*/
end
program define replic   /*more work needed!!*/

        local intercept_m `1'
        local intercept_v `2'
        local slope `3'
        local prevelence_x `4'
        local nstudies  `5'
        local senario `6'
forvalues replic=1(1)$repl { /* replication loop start from here*/

        forvalues j= 1(1)`nstudies'     {                       /*how mnay simulated IPD_i in each meta analysis*/
                *call onestudy for IPD
                IPD_j `intercept_m' `intercept_v' `slope' `prevelence_x'  /*code to simulate the Y and X using program
IPD_j*/

                gen study_id= `j'  /*ID for each IPD_J data within one replication*/

                tempfile ipd_`replic'_`j'     /*create a temporary file and save for each IPD_j in each replication*/

                        if `j'>1 {
                                        local j_former =`j'-1
                                        append using "C:\my thesis\middata\senario_`senario'ipd_`j_former'"
/*append last saved file from study_2 within same replication*/
                                                erase
"C:\my thesis\middata\senario_`senario'ipd_`j_former'.dta" /*remove the appended file*/
                                                }/*for the */

                        *save "C:\my thesis\middata\senario_`senario'ipd_`replic'_`j'", replace /*save all data,
removed*/

                        save "C:\my thesis\middata\senario_`senario'ipd_`j'", replace

                        }                               /*end of loop for each meta analysis*
```

```
                        *****keep data for checking purpose*****
                        erase "C:\my thesis\middata\senario_`senario'ipd_`nstudies'.dta"

                gen replic_id = `replic' /*create ID variable for each replication*/

                tempfile repl_`replic'
                if `replic'>1 {
                                        local repl_former =`replic'-1
                                        append using
"C:\my thesis\middata\senario_`senario'repl_`repl_former'"  /*append last saved file*/
                                                erase
"C:\my thesis\middata\senario_`senario'repl_`repl_former'.dta"
                                                }/*for the */
                        save "C:\my thesis\middata\senario_`senario'repl_`replic'" , replace
                        *save replic_`replic', replace
                        /*for code checking*/
                        if `replic'==$repl {
                                save "C:\my thesis\data\senario_`senario'", replace
                                                } /*save the last file for permernant: senario 1 all replication*/

                                        } /* end replication*/


                                        *****keep data for checking purpose*****

                *erase "C:\my thesis\middata\senario_`senario'repl_$repl.dta"         /*remove the last repl file*/
end
program two_step
        local intercept_m `1'
        local intercept_v `2'
        local slope `3'
        local prevelence_x `4'
        local nstudies `5'
        local senario `6'
        capture log close
        capture log using "C:\my thesis\two_step\cat_twostp_senario_`senario'", replace
        tempname step2_result  /*record all parameter and beta, se_beta value for each senario of each primary
study*/
        capture postfile `step2_result' senario intercept_m intercept_v slope lymphnod n_studies  beta se_beta  study
replic using "C:\my thesis\two_step\two_step_result_senario`senario'", replace
        forvalues replic = 1(1)$repl {    /*must change replic*/
                forvalues ipd_j = 1(1) `nstudies' {
                 capture quietly logit y x if replic_id == `replic'& study_id == `ipd_j'

                        if _rc==0 {
                                        local b_x = _b[x]
                                        local se_x = _se[x]
                                                }
                        if _rc>0 {
                                di "problemb data for 2stp is scenerio= " `senario' "    replication= " `replic' "
studyID= " `ipd_j'
                                        local b_x = .
                                        local se_x = .
                                                }
        capture post `step2_result' (`senario') (`intercept_m') (`intercept_v') (`slope') (`prevelence_x') (`nstudies')
(`b_x') (`se_x') (`ipd_j') (`replic')  /*save parameter setting for each senario*/
                                        }  /*each meta-analysis loop*/
                        } /*replication loop*/
        capture postclose `step2_result'  /*close to save senario parameter*/
                *** starting meta analysis beta for each senario***
        *reading two-step results for each senario
        use "C:\my thesis\two_step\two_step_result_senario`senario'"         , clear
        drop if se_beta==0 |se_beta==. /*remove no variance data in Logit model*/
```

```
	tempname step2_meta  /*record all parameter and beta, se_beta value for each senario of each primary study*/
	 postfile `step2_meta' senario intercept_m intercept_v slope lymphnod n_studies  beta_mt se_mt replic using "C:\my thesis\two_step\two_step_meta_senario`senario'", replace

			for values replication = 1(1)$repl {
			capture quietly metan  beta se_beta if replic==`replication', nograph notable
				capture post `step2_meta'  (`senario')  (`intercept_m')  (`intercept_v')  (`slope') (`prevelence_x') (`nstudies')  (`r(ES)') (`r(seES)') (`replication')
					}
	postclose `step2_meta'
	*** finish meta anlysis for senario and save file in two_step_meta_`senario'***
			capture log close
	end
program step_cluster
	local intercept_m `1'
	local intercept_v `2'
	local slope `3'
	local prevelence_x `4'
	local nstudies  `5'
	local senario `6'

	capture log close
	capture log using "C:\my thesis\step_cluster\cat_cluster_senario_`senario'", replace
	tempname cluster_result  /*record all parameter and beta, se_beta value for each senario of each primary study*/
	postfile `cluster_result' senario intercept_m intercept_v slope lymphnod n_studies  beta_cls se_cls replic using "C:\my thesis\step_cluster\cluster_result_senario`senario'", replace

	forvalues replic = 1(1)$repl {    /*must change replic*/

			capture quietly logit y x if replic_id == `replic'
				if _rc==0 {
					local b_x = _b[x]
					local se_x = _se[x]
				}

				if _rc>0 {
					di "problemb data for cluster is scenerio= " `senario' "   replication= " `replic'
					local b_x = .
					local se_x = .
						}
	capture  post  `cluster_result'  (`senario')  (`intercept_m')  (`intercept_v')  (`slope')  (`prevelence_x')  (`nstudies') (`b_x') (`se_x') (`replic') /*save parameter setting for each senario*/
					}  /*each meta-analysis loop*/
	capture postclose `cluster_result'  /*close to save senario parameter*/
	capture log close
end
program step_dummy
	local intercept_m `1'
	local intercept_v `2'
	local slope `3'
	local prevelence_x `4'
	local nstudies  `5'
	local senario `6'
	capture log close
	capture log using "C:\my thesis\step_dummy\cat_dummy_senario_`senario'", replace
	tempname dummy_result  /*record all parameter and beta, se_beta value for each senario of each primary study*/
	postfile `dummy_result' senario intercept_m intercept_v slope lymphnod n_studies  beta_dmy se_dmy replic using "C:\my thesis\step_dummy\dummy_result_senario`senario'", replace

	forvalues replic = 1(1)$repl {    /*must change replic*/
```

```
                capture quietly xi: logit y x i.study_id if replic_id == `replic'
                        if _rc==0 {
                                local b_x = _b[x]
                                local se_x = _se[x]
                        }
                if _rc>0 {
                        di "problemb data for DUMMY  is scenerio= " `senario' "   replication= " `replic'
                                local b_x = .
                                local se_x = .
                                }
        capture post `dummy_result' (`senario') (`intercept_m') (`intercept_v') (`slope') (`prevelence_x') (`nstudies')
(`b_x') (`se_x') (`replic') /*save parameter setting for each senario*/
                                }  /*each meta-analysis loop*/
        capture postclose `dummy_result'  /*close to save senario parameter*/
        capture log close
end

*** conbine 3 results from one senario data
program conbin_result
        local intercept_m `1'
        local intercept_v `2'
        local slope `3'
        local prevelence_x `4'
        local nstudies  `5'
        local senario `6'

use "C:\my thesis\two_step\two_step_meta_senario`senario'", clear
merge 1:1 senario intercept_m intercept_v slope lymphnod n_studies replic using "C:\my
thesis\step_cluster\cluster_result_senario`senario'", generate(frm_cls)
merge 1:1 senario intercept_m intercept_v slope lymphnod n_studies replic using "C:\my
thesis\step_dummy\dummy_result_senario`senario'",
save  "C:\my thesis\result\cat_result_senario_`senario'", replace   /*file name as categorical x*/
end
```

Appendix D2: The simulation STATA code program for continuous PF.

```
* Jan-21-2011
*C:\my thesis
*need cal macro in file gadasim first
*continous covariate code
do C:\my thesis\code\Ghada_sim_cont
** parameter seting from here
clear
clear matrix
set memory 1000m
set more off
set seed 12345678
cd "C:\my thesis"
tempname simparm   /*record all parameter and population value for each senario*/
postfile `simparm' senario inercept_m inercept_v slope age_m age_sd n_studies using C:\my thesis\senario_monitor.dta,
replace
*****set different parameter here, following are Gadha's code
local a_m  "-2.10"      /*intercept mean*/
local a_v  "1.5"      /*intercept SD*/
local b  "0.30 0.1 0"   /*treatment*/
local m_age  "4"  /* prevenlance*/
local sd_age "1.5"
local n_studies "5" /* number of IPD_j in each meta analysis */

global repl = 1000   /*decide how many replication used*/
****** DO not change following code******
local senario = 1
foreach intpt_m in `a_m' {
        foreach intpt_v in `a_v'  {
                foreach slp in `b'                  {
                        foreach age_m in `m_age' {
                                foreach age_sd in `sd_age' {
                                        foreach n_meta in `n_studies' {

                                quietly replic `intpt_m' `intpt_v' `slp' `age_m' `age_sd' `n_meta'  `senario'
/*simulate data*/

                                post `simparm' (`senario') (`intpt_m') (`intpt_v') (`slp') (`age_m') (`age_sd')
(`n_meta') /*save parameter setting for each senario*/
                                *Analysis data here or seperately?
                                use "C:\my thesis\data\senario_`senario'", clear
                                two_step `intpt_m' `intpt_v' `slp' `age_m' `age_sd' `n_meta'  `senario' /*runing
two_step analysis*/
                                use "C:\my thesis\data\senario_`senario'", clear
                                step_cluster `intpt_m' `intpt_v' `slp' `age_m' `age_sd' `n_meta'  `senario' /*runing
one_step cluster*/
                                use "C:\my thesis\data\senario_`senario'", clear
                                step_dummy `intpt_m' `intpt_v' `slp' `age_m' `age_sd' `n_meta'  `senario'
/*runing one_step dummy*/
                                conbin_result `intpt_m' `intpt_v' `slp' `age_m' `age_sd' `n_meta'  `senario'
/*combine 3 kind of result together for same senario*/
                                local ++senario
                                                } /*end of study number loop*/
                                        } /*end of age sd loop*/
                                } /*end of age mean loop*/
                        } /*end of slope loop*/
                } /*end of  alpha variance loop*/
```

```
	} /*end of  intercept loop*/

	postclose `simparm'  /*close to save senario parameter*/

	***********
**Jan21-2011
* programe here are called by file My thesis_cont
* programm are
*IPD_j
*replic
*two_step
*step_cluster
*step_dummy
*continous covriate
**********************
capture program drop IPD_j
capture program drop replic
capture program drop two_step
capture program drop step_cluster
capture program drop step_dummy
capture program drop conbin_result

program define IPD_j

	local alpha_m `1'
	local alpha_v `2'
	local beta `3'
	local p_x `4'
	local p_xsd `5'
		clear /*clear the file in memory for new simulation*/
	local alpha= (rnormal(`alpha_m', `alpha_v'))
		local n_ipd_j = int(30+(runiform()*(1000-30))) /*30~100 subjects for each primary study_j gadha*/
	quietly set obs `n_ipd_j' /*how many people in each IPD study*/

	gen x = (rnormal(`p_x', `p_xsd')) /*x is age, centered X, doesn't affect beta and intercept is alpha!!*/
	gen y = runiform() < invlogit(`alpha'+ `beta'*x) /*Thompson code*/
end
program define replic   /*more work needed!!*/
	local intercept_m `1'
	local intercept_v `2'
	local slope `3'
	local age_m `4'
	local age_sd `5'
	local nstudies  `6'
	local senario `7'
	forvalues replic=1(1)$repl { /* replication loop start from here*/
	forvalues j= 1(1)`nstudies'     {                       /*how mnay simulated IPD_i in each meta analysis*/
		*call onestudy for IPD
		IPD_j `intercept_m' `intercept_v' `slope' `age_m' `age_sd'   /*code to simulate the Y and X using
program IPD_j*/

		gen study_id= `j'  /*ID for each IPD_J data within one replication*/

		tempfile ipd_`replic'_`j'     /*create a temporary file and save for each IPD_j in each replication*/

			if `j'>1 {
					local j_former =`j'-1
					append  using  "C:\my  thesis\middata\senario_`senario'ipd_`j_former'"
/*append last saved file from study_2 within same replication*/

						erase
"C:\my thesis\middata\senario_`senario'ipd_`j_former'.dta" /*remove the appended file*/
						}/*for the */
```

```
                        *save "C:\my thesis\middata\senario_`senario'ipd_`replic'_`j'", replace /*save all data, removed*/

                        save "C:\my thesis\middata\senario_`senario'ipd_`j'", replace

                        }                                   /*end of loop for each meta analysis*/

                                                            ****keep data for checking purpose
                        erase "C:\my thesis\middata\senario_`senario'ipd_`nstudies'.dta"

                gen replic_id = `replic' /*create ID variable for each replication*/

                tempfile repl_`replic'
                if `replic'>1 {
                                                    local repl_former =`replic'-1
                                                    append using
"C:\my thesis\middata\senario_`senario'repl_`repl_former'"  /*append last saved file*/
                                                      erase
"C:\my thesis\middata\senario_`senario'repl_`repl_former'.dta"
                                                    }/*for the */
                        save "C:\my thesis\middata\senario_`senario'repl_`replic'" , replace
                        *save replic_`replic', replace
                        /*for code checking*/

                        if `replic'==$repl {

                                save "C:\my thesis\data\senario_`senario'", replace

                                                    } /*save the last file for permernant: senario 1 all replication*/


                                                    } /* end replication*/

                ****keep data for checking purpose
                *erase "C:\my thesis\middata\senario_`senario'repl_$repl.dta"          /*remove the last repl file*/

end
program two_step
        local intercept_m `1'
        local intercept_v `2'
        local slope `3'
        local age_m `4'
        local age_sd `5'
        local nstudies  `6'
        local senario `7'

        capture log close
        capture log using "C:\my thesis\two_step\twostp_senario_`senario'", replace

        tempname step2_result  /*record all parameter and beta, se_beta value for each senario of each primary study*/
        capture postfile `step2_result' senario intercept_m intercept_v slope age_m age_sd n_studies  beta se_beta study replic using "C:\my thesis\two_step\two_step_result_senario`senario'", replace

        forvalues replic = 1(1)$repl {    /*must change replic*/
                forvalues ipd_j = 1(1) `nstudies' {

                        capture quietly logit y x if replic_id == `replic'& study_id == `ipd_j'

                                if _rc==0 {

                                        local b_x = _b[x]
```

```
                                        local se_x = _se[x]
                                                                        }
                                    /*report problem dataset*/
                                    if _rc>0 {
                                                di "problem data for 2stp is scenerio= " `senario' "
replication= " `replic' " studyID= " `ipd_j'
                                        local b_x = .
                                        local se_x = .
                                        }

                            capture post `step2_result' (`senario') (`intercept_m') (`intercept_v') (`slope') (`age_m')
(`age_sd') (`nstudies') (`b_x') (`se_x') (`ipd_j') (`replic')  /*save parameter setting for each senario*/
                                        }  /*each meta-analysis loop*/
                            } /*replication loop*/

        capture postclose `step2_result'  /*close to save senario parameter*/

        *** starting meta analysis beta for each senario***
        *reading two-step results for each senario
        use "C:\my thesis\two_step\two_step_result_senario`senario'"        , clear
        drop if se_beta==0 |se_beta==. /*remove no variance data in Logit model*/
        tempname step2_meta  /*record all parameter and beta, se_beta value for each senario of each primary
study*/
        capture postfile `step2_meta' senario intercept_m intercept_v slope age_m age_sd n_studies  beta_mt se_mt
replic using "C:\my thesis\two_step\two_step_meta_senario`senario'", replace

                forvalues replication = 1(1)$repl {
                        capture quietly metan  beta se_beta if replic==`replication', nograph notable

                        capture post `step2_meta'    (`senario') (`intercept_m') (`intercept_v') (`slope') (`age_m')
(`age_sd') (`nstudies')  (`r(ES)') (`r(seES)') (`replication')
                                }
        postclose `step2_meta'
        *** finish meta anlysis for senario and save file in two_step_meta_`senario'***

        capture log close
end

program step_cluster
        local intercept_m `1'
        local intercept_v `2'
        local slope `3'
        local age_m `4'
        local age_sd `5'
        local nstudies  `6'
        local senario `7'

        capture log close
        capture log using "C:\my thesis\step_cluster\cluster_senario_`senario'", replace

        tempname cluster_result  /*record all parameter and beta, se_beta value for each senario of each primary
study*/
        capture postfile `cluster_result' senario intercept_m intercept_v slope age_m age_sd n_studies  beta_cls se_cls
replic using "C:\my thesis\step_cluster\cluster_result_senario`senario'", replace

        forvalues replic = 1(1)$repl {    /*must change replic*/

                capture quietly logit y x if replic_id == `replic'
                                if _rc==0 {

                                        local b_x = _b[x]
                                        local se_x = _se[x]
```

```
                                        }
                /*report problem dataset*/
                                if _rc>0 {
                                        di "problem data for CLUSTER is scenerio= " `senario' "   replication= "
`replic'
                                        local b_x = .
                                        local se_x = .
                                        }

                capture post `cluster_result' (`senario') (`intercept_m') (`intercept_v') (`slope') (`age_m') (`age_sd')
(`nstudies') (`b_x') (`se_x') (`replic') /*save parameter setting for each senario*/
                                        }  /*each meta-analysis loop*/
        capture postclose `cluster_result'  /*close to save senario parameter*/
        capture log close

end
program step_dummy
        local intercept_m `1'
        local intercept_v `2'
        local slope `3'
        local age_m `4'
        local age_sd `5'
        local nstudies  `6'
        local senario `7'

        capture log close
        capture log using "C:\my thesis\step_dummy\dummy_senario_`senario'", replace

        tempname dummy_result  /*record all parameter and beta, se_beta value for each senario of each primary
study*/
        capture postfile `dummy_result' senario intercept_m intercept_v slope age_m age_sd n_studies  beta_dmy
se_dmy replic using "C:\my thesis\step_dummy\dummy_result_senario`senario'", replace

        forvalues replic = 1(1)$repl {   /*must change replic*/

                capture quietly xi: logit y x i.study_id if replic_id == `replic'

                                        if _rc==0 {

                                        local b_x = _b[x]
                                        local se_x = _se[x]

                                        }
                /*report problem dataset*/
                                if _rc>0 {
                                        di "problem data for DUMMY is scenerio= " `senario' "   replication= "
`replic'
                                        local b_x = .
                                        local se_x = .
                                        }

                post  `dummy_result' (`senario') (`intercept_m') (`intercept_v') (`slope') (`age_m') (`age_sd')
(`nstudies') (`b_x') (`se_x') (`replic') /*save parameter setting for each senario*/
                                        }  /*each meta-analysis loop*/
                postclose `dummy_result'  /*close to save senario parameter*/

        capture log close
        end

*** conbine 3 results from one senario data
rogram conbin_result
        local intercept_m `1'
        local intercept_v `2'
```

```
        local slope `3'
        local age_m `4'
        local age_sd `5'
        local nstudies  `6'
        local senario `7'
use "C:\my thesis\two_step\two_step_meta_senario`senario'", clear
merge 1:1 senario intercept_m intercept_v slope age_m age_sd n_studies replic using "C:\my thesis\step_cluster\cluster_result_senario`senario'", generate(frm_cls)
merge 1:1 senario intercept_m intercept_v slope age_m age_sd n_studies replic using "C:\my thesis\step_dummy\dummy_result_senario`senario'",
save  "C:\my thesis\result\result_senario_`senario'", replace
/*remove mid-process files*/
 End
```

Appendix D3: Create 95% CIs and how many of the 1000 simulations give a CI containing true value β

```
gen lower_mt = beta_mt - (1.96*se_mt)
gen upper_mt = beta_mt + (1.96*se_mt)
gen included_mt = 0
replace included_mt = 1 if  lower_mt <= β & upper_mt >= β
gen lower_cls = beta_cls - (1.96*se_cls)
gen upper_cls = beta_cls + (1.96*se_cls)
gen included_cls = 0
replace included_cls = 1 if  lower_cls <= β & upper_cls >= β
gen lower_dmy = beta_dmy - (1.96*se_dmy)
gen upper_dmy = beta_dmy + (1.96*se_dmy)
gen included_dmy = 0
replace included_dmy = 1 if  lower_dmy <= 0 & upper_dmy >= 0
total included_mt included_cls included_dmy
```

Appendix D4: Simulation results for n=10 studies for binary PF

Table D4.1: Simulation results for the pooled effect size,$\widehat{\boldsymbol{\beta}}$, for three models of IPD meta-analysis,(two-step, one-step ignoring clustering and one-step including indicator variable), with prevalence =0.5and the sample size with-in each study is between 30 to 1000 observations, with the number of studies n=10 and 1000 simulations, the true values of the pooled effect size,$\boldsymbol{\beta}$, is shown in the table; the standard error for $\propto$ is 0 and 0.25.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (s. e. of $\propto$)	*prevalence*	*True* β	Mean of $\widehat{\beta}$	Bias	percentage bias	MSE of β	Percentage coverage	Mean s.e. of $\widehat{\beta}$
1	10	Model 2	Two-step	-1.27(0)	0.5	0.9	0.90	0.00	-0.49	0.01	95.00	0.06
		Model 6	Ignoring clustering	-1.27(0)	0.5	0.9	0.89	-0.01	-1.48	0.01	94.30	0.06
		Model 8	One-step including indicator variable.	-1.27(0)	0.5	0.9	0.90	0.00	-0.14	0.01	94.90	0.06
2	10	Model 2	Two-step	-1.27(0)	0.5	0.1	0.10	0.00	1.28	0.01	95.00	0.10
		Model 6	Ignoring clustering	-1.27(0)	0.5	0.1	0.10	0.00	0.62	0.01	94.70	0.10
		Model 8	One-step including indicator variable.	-1.27(0)	0.5	0.1	0.10	0.00	1.80	0.01	94.70	0.10
3	10	Model 2	Two-step	-1.27(0)	0.5	0	0	0.00	NA	0.01	95.10	0.07
		Model 6	Ignoring clustering	-1.27(0)	0.5	0	0	0.00	NA	0.01	94.70	0.07
		Model 8	One-step including indicator variable.	-1.27(0)	0.5	0	0	0.00	NA	0.01	94.90	0.07
4	10	Model 2	Two-step	-1.27(0.25)	0.5	0.9	0.90	0.00	-0.17	0.01	95.10	0.09
		Model 6	Ignoring clustering	-1.27(0.25)	0.5	0.9	0.90	0.00	-0.11	0.01	95.20	0.09
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.5	0.9	0.90	0.00	0.07	0.01	95.40	0.09
5	10	Model 2	Two-step	-1.27(0.25)	0.5	0.1	0.10	0.00	-2.65	0.01	95.40	0.06
		Model 6	Ignoring clustering	-1.27(0.25)	0.5	0.1	0.10	0.00	-2.63	0.01	95.50	0.06
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.5	0.1	0.10	0.00	-2.48	0.01	95.30	0.06
6	10	Model 2	Two-step	-1.27(0.25)	0.5	0	0	0.00	NA	0.01	95.50	0.06
		Model 6	Ignoring clustering	-1.27(0.25)	0.5	0	0	0.00	NA	0.01	95.50	0.06
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.5	0	0	0.00	NA	0.01	95.40	0.06

Table D4.2: Simulation results for the pooled effect size, $\hat{\beta}$, for three models of IPD meta-analyses (two-step, one-step ignoring clustering and one-step including indicator variable), with prevalence=0.2 and the sample size with-in each study is between 30 to 1000 observations, with the number of studies n=10 and 1000 simulations, the true values of the pooled effect size,β, is shown in the table; the standard error for $\propto$ is 0 and 0.25.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (s. e of $\propto$)	Prevalence	True β	Mean of $\widehat{\beta}$	Bias	percentage bias	MSE of β	Percentage coverage	Mean s.e. of $\widehat{\beta}$
7	10	Model 2	Two-step	-1.27(0)	0.2	0.9	0.89	-0.01	-0.63	0.01	94.80	0.08
		Model 6	Ignoring clustering	-1.27(0)	0.2	0.9	0.90	0.00	0.12	0.01	94.30	0.08
		Model 8	One-step including indicator variable.	-1.27(0)	0.2	0.9	0.90	0.00	0.29	0.01	94.40	0.08
8	10	Model 2	Two-step	-1.27(0)	0.2	0.1	0.10	0.00	-2.52	0.01	95.90	0.07
		Model 6	Ignoring clustering	-1.27(0)	0.2	0.1	0.10	0.00	-1.87	0.01	95.60	0.07
		Model 8	One-step including indicator variable.	-1.27(0)	0.2	0.1	0.10	0.00	-1.76	0.01	95.60	0.07
9	10	Model 2	Two-step	-1.27(0)	0.2	0	0.00	0.00	NA	0.01	95.00	0.07
		Model 6	Ignoring clustering	-1.27(0)	0.2	0	0.00	0.00	NA	0.01	94.80	0.07
		Model 8	One-step including indicator variable.	-1.27(0)	0.2	0	0.00	0.00	NA	0.01	94.80	0.07
10	10	Model 2	Two-step	-1.27(0.25)	0.2	0.9	0.90	0.00	0.30	0.01	95.20	0.08
		Model 6	Ignoring clustering	-1.27(0.25)	0.2	0.9	0.90	0.00	-1.18	0.01	94.20	0.08
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.2	0.9	0.90	0.00	0.11	0.01	94.90	0.08
11	10	Model 2	Two-step	-1.27(0.25)	0.2	0.1	0.11	0.01	10.45	0.01	95.30	0.11
		Model 6	Ignoring clustering	-1.27(0.25)	0.2	0.1	0.10	0.00	-0.52	0.01	95.20	0.10
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.2	0.1	0.10	0.00	0.39	0.01	95.40	0.10
10	10	Model 2	Two-step	-1.27(0.25)	0.2	0	0.01	0.01	NA	0.01	95.50	0.09
		Model 6	Ignoring clustering	-1.27(0.25)	0.2	0	0.00	0.00	NA	0.01	94.70	0.09
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.2	0	0.00	0.00	NA	0.01	95.20	0.09

Table D4.3: Simulation results for the pooled effect size, $\boldsymbol{\beta}$, for three models of IPD meta-analyses (two-step, one-step ignoring clustering and one-step including indicator variable, with prevalence =0.2 and the sample size with-in each study is between 30 to 1000 observations, with the number of studies n=10 and 1000 simulations, the standard error for $\propto$ is 1.5.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (s.e. of $\propto$)	Prevalence	True $\boldsymbol{\beta}$	Mean of $\widehat{\boldsymbol{\beta}}$	Bias	percentage bias	MSE of $\boldsymbol{\beta}$	Percentage coverage	Mean s.e. of $\widehat{\boldsymbol{\beta}}$
13	10	Model 2	Two-step	-1.27(1.5)	0.2	0.9	0.90	0.00	-0.33	0.01	95.90	0.09
		Model 6	Ignoring clustering	-1.27(1.5)	0.2	0.9	0.67	-0.23	-25.57	0.07	23.50	0.07
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.2	0.9	0.90	0.00	-0.23	0.01	96.10	0.09
14	10	Model 2	Two-step	-1.27(1.5)	0.2	0.1	0.11	0.01	10.39	0.01	94.50	0.09
		Model 6	Ignoring clustering	-1.27(1.5)	0.2	0.1	0.08	-0.02	-23.89	0.01	93.70	0.08
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.2	0.1	0.10	0.00	0.29	0.01	94.00	0.09
15	10	Model 2	Two-step	-1.27(1.5)	0.2	0	0.01	0.01	NA	0.02	94.70	0.09
		Model 6	Ignoring clustering	-1.27(1.5)	0.2	0	0.00	0.00	NA	0.02	94.40	0.08
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.2	0	-0.01	-0.01	NA	0.02	94.20	0.09
16	10	Model 2	Two-step	-1.27(1.5)	0.5	0.9	0.86	-0.04	-4.44	0.04	96.70	0.21
		Model 6	Ignoring clustering	-1.27(1.5)	0.5	0.9	0.67	-0.23	-25.56	0.09	71.80	0.17
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.5	0.9	0.91	0.01	1.11	0.04	96.10	0.21
17	10	Model 2	Two-step	-1.27(1.5)	0.5	0.1	0.10	0.00	0.00	0.04	96.3	0.22
		Model 6	Ignoring clustering	-1.27(1.5)	0.5	0.1	0.08	-0.02	-20.00	0.03	95.6	0.18
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.5	0.1	0.11	0.01	10.00	0.04	94.7	0.21
18	10	Model 2	Two-step	-1.27(1.5)	0.5	0	-0.01	-0.01	NA	0.04	96.50	0.22
		Model 6	Ignoring clustering	-1.27(1.5)	0.5	0	0.00	0.00	NA	0.03	95.70	0.18
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.5	0	-0.01	-0.01	NA	0.04	95.20	0.21

Table D4.4: Simulation results for the pooled effect size,$\widehat{\boldsymbol{\beta}}$, for three models of IPD meta-analysis, two-step, one-step ignoring clustering and one-step including indicator variable, with prevalence =0.5and the sample size with-in each study is between 30 to 100 observations, with the number of studies n=10 and 1000 simulations, the true values of the pooled effect size,$\boldsymbol{\beta}$, is shown in the table; the standard error for $\propto$ is 0 and 0.25.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (s. e. of $\propto$)	*prevalence*	Mean of $\widehat{\boldsymbol{\beta}}$	$\boldsymbol{\beta}$	Bias	percentage bias	MSE of $\boldsymbol{\beta}$	Percentage coverage	Mean s.e. of $\widehat{\boldsymbol{\beta}}$
1	10	Model 2	Two-step	-1.27(0)	0.5	0.90	0.9	0.00	-0.18	0.03	95.90	0.17
		Model 6	Ignoring clustering	-1.27(0)	0.5	0.91	0.9	0.01	0.50	0.03	95.90	0.17
		Model 8	One-step including indicator variable.	-1.27(0)	0.5	0.92	0.9	0.02	1.97	0.03	95.00	0.17
2	10	Model 2	Two-step	-1.27(0)	0.5	0.10	0.10	0.00	3.73	0.03	95.9	0.16
		Model 6	Ignoring clustering	-1.27(0)	0.5	0.11	0.10	0.01	4.95	0.03	95.7	0.16
		Model 8	One-step including indicator variable.	-1.27(0)	0.5	0.11	0.10	0.01	5.90	0.03	95.5	0.16
3	10	Model 2	Two-step	-1.27(0)	0.5	0.00	0	0.00	NA	0.03	96.50	0.16
		Model 6	Ignoring clustering	-1.27(0)	0.5	0.00	0	0.00	NA	0.03	96.70	0.16
		Model 8	One-step including indicator variable.	-1.27(0)	0.5	0.00	0	0.00	NA	0.03	95.80	0.16
4	10	Model 2	Two-step	-1.27(0.25)	0.5	0.88	0.9	-0.03	-2.77	0.03	96.30	0.17
		Model 6	Ignoring clustering	-1.27(0.25)	0.5	0.89	0.9	-0.01	-1.57	0.03	95.50	0.18
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.5	0.91	0.9	0.01	0.85	0.03	95.30	0.18
5	10	Model 2	Two-step	-1.27(0.25)	0.5	0.10	0.10	0.00	-2.45	0.04	95.90	0.20
		Model 6	Ignoring clustering	-1.27(0.25)	0.5	0.10	0.10	0.00	1.25	0.04	95.10	0.19
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.5	0.10	0.10	0.00	3.19	0.04	94.50	0.19
6	10	Model 2	Two-step	-1.27(0.25)	0.5	-0.01	0	-0.01	NA	0.04	96.70	0.20
		Model 6	Ignoring clustering	-1.27(0.25)	0.5	-0.01	0	-0.01	NA	0.04	94.90	0.19
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.5	0.01	0	0.01	NA	0.04	95.20	0.20

Table D4.5: Simulation results for the pooled effect size, $\widehat{\boldsymbol{\beta}}$, for three models of IPD meta-analyses (two-step, one-step ignoring clustering and

one-step including indicator variable, with prevalence=0.2 and the sample size with-in each study is between 30 to 100 observations, with the number of studies n=10 and 1000 simulations, the true values of the pooled effect size,$\boldsymbol{\beta}$, is shown in the table; the standard error for $\propto$ is 0 and 0.25.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (s.e of $\propto$)	Prevalence	Mean of $\widehat{\boldsymbol{\beta}}$	True $\boldsymbol{\beta}$	Bias	percentage bias	MSE of $\boldsymbol{\beta}$	Percentage coverage	Mean s.e. of $\widehat{\boldsymbol{\beta}}$
7	10	Model 2	Two-step	-1.27(0)	0.2	0.89	0.9	-0.01	-0.63	-7.15	94.80	0.08
		Model 6	Ignoring clustering	-1.27(0)	0.2	0.90	0.9	0.00	0.12	1.36	94.30	0.08
		Model 8	One-step including indicator variable.	-1.27(0)	0.2	0.90	0.9	0.00	0.29	3.24	94.40	0.08
8	10	Model 2	Two-step	-1.27(0)	0.2	0.10	0.10	0.00	-2.52	-3.63	95.90	0.07
		Model 6	Ignoring clustering	-1.27(0)	0.2	0.10	0.10	0.00	-1.87	-2.68	95.60	0.07
		Model 8	One-step including indicator variable.	-1.27(0)	0.2	0.10	0.10	0.00	-1.76	-2.52	95.60	0.07
9	10	Model 2	Two-step	-1.27(0)	0.2	0.00	0	0.00	NA	-0.52	95.00	0.07
		Model 6	Ignoring clustering	-1.27(0)	0.2	0.00	0	0.00	NA	-0.26	94.80	0.07
		Model 8	One-step including indicator variable.	-1.27(0)	0.2	0.00	0	0.00	NA	-3.17	94.80	0.07
10	10	Model 2	Two-step	-1.27(0.25)	0.2	0.90	0.9	0.00	0.30	0.01	95.20	0.08
		Model 6	Ignoring clustering	-1.27(0.25)	0.2	0.90	0.9	0.00	-1.18	0.01	94.20	0.08
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.2	0.90	0.9	0.00	0.11	0.01	94.90	0.08
11	10	Model 2	Two-step	-1.27(0.25)	0.2	0.11	0.1	0.01	10.45	0.01	95.30	0.11
		Model 6	Ignoring clustering	-1.27(0.25)	0.2	0.10	0.1	0.00	-0.52	0.01	95.20	0.10
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.2	0.10	0.1	0.00	0.39	0.01	95.40	0.10
10	10	Model 2	Two-step	-1.27(0.25)	0.2	0.01	0	0.01	NA	0.01	95.50	0.09
		Model 6	Ignoring clustering	-1.27(0.25)	0.2	0.00	0	0.00	NA	0.01	94.70	0.09
		Model 8	One-step including indicator variable.	-1.27(0.25)	0.2	0.00	0	0.00	NA	0.01	95.20	0.09

Table D4.6: Simulation results for the pooled effect size, for three models of IPD meta-analyses (two-step, one-step ignoring clustering and one-step including indicator variable, with prevalence =0.2 and the sample size with-in each study is between 30 to 100 observations, with the number of studies n=10 and 1000 simulations, the standard error for $\propto$ is 1.5.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (s.e. of $\propto$)	Prevalence	Mean of $\widehat{\beta}$	True β	Bias	percentage bias	MSE of β	Percentage coverage	Mean s.e. of $\widehat{\beta}$
13	10	Model 2	Two-step	-1.27(1.5)	0.2	0.91	0.9	0.01	0.88	0.06	97.20	0.26
		Model 6	Ignoring clustering	-1.27(1.5)	0.2	0.67	0.9	-0.23	-25.65	0.10	78.50	0.21
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.2	0.92	0.9	0.02	2.06	0.06	95.10	0.25
14	10	Model 2	Two-step	-1.27(1.5)	0.2	0.19	0.1	0.09	91.34	0.07	95.80	0.28
		Model 6	Ignoring clustering	-1.27(1.5)	0.2	0.08	0.1	-0.02	-22.30	0.05	95.60	0.22
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.2	0.10	0.1	0.00	3.29	0.07	95.20	0.26
15	10	Model 2	Two-step	-1.27(1.5)	0.2	0.10	0	0.10	NA	0.08	95.60	0.28
		Model 6	Ignoring clustering	-1.27(1.5)	0.2	0.00	0	0.00	NA	0.05	95.00	0.23
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.2	0.00	0	0.00	NA	0.07	95.10	0.27
16	10	Model 2	Two-step	-1.27(1.5)	0.5	0.89	0.9	-0.01	1.11	0.01	94.00	0.07
		Model 6	Ignoring clustering	-1.27(1.5)	0.5	0.67	0.9	-0.23	-25.56	0.06	14.80	0.06
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.5	0.90	0.9	0.00	0.00	0.00	93.90	0.07
17	10	Model 2	Two-step	-1.27(1.5)	0.5	0.10	0.1	0.00	0.00	0.00	96.00	0.07
		Model 6	Ignoring clustering	-1.27(1.5)	0.5	0.07	0.1	-0.03	-30.00	0.01	93.00	0.06
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.5	0.10	0.1	0.00	0.00	0.01	95.90	0.07
18	10	Model 2	Two-step	-1.27(1.5)	0.5	0.00	0	0.00	NA	0.01	94.10	0.07
		Model 6	Ignoring clustering	-1.27(1.5)	0.5	0.00	0	0.00	NA	0.00	94.70	0.06
		Model 8	One-step including indicator variable.	-1.27(1.5)	0.5	0.00	0	0.00	NA	0.01	94.00	0.07

Appendix D5: Simulation results for n=10 studies for continuous PF

Table D5.1: Simulation results for the pooled effect size,$\widehat{\boldsymbol{\beta}}$, for three models of IPD meta-analysis, two-step, one-step ignoring clustering and one-step including indicator variable, mean age is 4 with standard deviation 1.5 and the sample size with-in each study is between 30 to 100 observations, with the number of studies n=10 and 100 simulations, the true values of the pooled effect size,$\boldsymbol{\beta}$, is shown in the table; the standard error for $\propto$ is 0.2 and 1.5.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (s. e. of $\propto$)	Mean of $\widehat{\boldsymbol{\beta}}$	$\boldsymbol{\beta}$	Bias	percentage bias	MSE of $\boldsymbol{\beta}$	Percentage coverage	Mean s.e. of $\widehat{\boldsymbol{\beta}}$
1	10	Model 2	Two-step	-2.1(0.2)	0.29	0.30	-0.006	-2.16	0.00	97.27	0.06
		Model 6	Ignoring clustering	-2.1(0.2)	0.30	0.30	0.000	-0.05	0.00	96.71	0.06
		Model 8	One-step including indicator variable.	-2.1(0.2)	0.31	0.30	0.007	2.25	0.00	96.85	0.06
2	10	Model 2	Two-step	-2.1(0.2)	0.10	0.10	0.002	1.68	0.00	95.59	0.04
		Model 6	Ignoring clustering	-2.1(0.2)	0.10	0.10	0.002	1.72	0.00	95.68	0.04
		Model 8	One-step including indicator variable.	-2.1(0.2)	0.10	0.10	0.002	2.26	0.00	95.41	0.04
3	10	Model 2	Two-step	-2.1(0.2)	0	0	0	NA	0.01	96.60	0.09
		Model 6	Ignoring clustering	-2.1(0.2)	0	0	0	NA	0.01	95.40	0.09
		Model 8	One-step including indicator variable.	-2.1(0.2)	0	0	0	NA	0.01	95.30	0.09
4	10	Model 2	Two-step	-2.1(1.5)	0.29	0.30	-0.01	-3.38	0.00	96.50	0.07
		Model 6	Ignoring clustering	-2.1(1.5)	0.22	0.30	-0.08	-25.57	0.01	72.10	0.06
		Model 8	One-step including indicator variable.	-2.1(1.5)	0.30	0.30	0.00	1.42	0.01	94.90	0.07
5	10	Model 2	Two-step	-2.1(1.5)	0.09	0.10	-0.01	-5.50	0.01	97.40	0.08
		Model 6	Ignoring clustering	-2.1(1.5)	0.08	0.10	-0.02	-24.37	0.00	94.80	0.07
		Model 8	One-step including indicator variable.	-2.1(1.5)	0.10	0.10	0.00	-0.53	0.01	95.70	0.08
6	10	Model 2	Two-step	-2.1(1.5)	0.00	0	0.00	NA	0.01	97.40	0.08
		Model 6	Ignoring clustering	-2.1(1.5)	0.00	0	0.00	NA	0.01	94.90	0.07
		Model 8	One-step including indicator variable.	-2.1(1.5)	0.00	0	0.00	NA	0.01	96.70	0.08

Table D5.2: Simulation results for the pooled effect size,$\widehat{\boldsymbol{\beta}}$, for three models of IPD meta-analysis, two-step, one-step ignoring clustering and one-step including indicator variable, mean age is 4 with standard deviation 1.5 and the sample size with-in each study is between 30 to 100 observations, with the number of studies n=10 and 1000 simulations, the true values of the pooled effect size,$\boldsymbol{\beta}$, is shown in the table; the standard error for $\propto$ is 0.2 and 1.5.

Scenarios	No. of studies(n)	Model	Meta-analysis model	$\propto$ (s. e. of $\propto$)	Mean of $\widehat{\boldsymbol{\beta}}$	$\boldsymbol{\beta}$	Bias	percentage bias	MSE of $\boldsymbol{\beta}$	Percentage coverage	Mean s.e. of $\widehat{\boldsymbol{\beta}}$
1	10	Model 2	Two-step	-2.1(0.2)	0.30	0.30	0.00	0.02	0.00	94.8	0.02
		Model 6	Ignoring clustering	-2.1(0.2)	0.30	0.30	0.00	-0.36	0.00	94.7	0.02
		Model 8	One-step including indicator variable.	-2.1(0.2)	0.30	0.30	0.00	0.56	0.00	94.4	0.02
2	10	Model 2	Two-step	-2.1(0.2)	0.10	0.10	0.00	0.77	0.00	95.8	0.03
		Model 6	Ignoring clustering	-2.1(0.2)	0.10	0.10	0.00	0.67	0.00	95.9	0.03
		Model 8	One-step including indicator variable.	-2.1(0.2)	0.10	0.10	0.00	1.36	0.00	95.8	0.03
3	10	Model 2	Two-step	-2.1(0.2)	0.00	0.00	0.00	NA	0.00	95.6	0.03
		Model 6	Ignoring clustering	-2.1(0.2)	0.00	0.00	0.00	NA	0.00	95.3	0.03
		Model 8	One-step including indicator variable.	-2.1(0.2)	0.00	0.00	0.00	NA	0.00	95.5	0.03
4	10	Model 2	Two-step	-2.1(1.5)	0.30	0.30	0.00	-0.28	0.00	94.10	0.02
		Model 6	Ignoring clustering	-2.1(1.5)	0.23	0.30	-0.07	-24.97	0.01	18.60	0.02
		Model 8	One-step including indicator variable.	-2.1(1.5)	0.30	0.30	0.00	0.27	0.00	93.30	0.02
5	10	Model 2	Two-step	-2.1(1.5)	0.10	0.10	0.00	-2.76	0.00	95.20	0.03
		Model 6	Ignoring clustering	-2.1(1.5)	0.08	0.10	-0.02	-23.75	0.00	79.10	0.02
		Model 8	One-step including indicator variable.	-2.1(1.5)	0.10	0.10	0.00	-2.18	0.00	95.20	0.03
6	10	Model 2	Two-step	-2.1(1.5)	0.00	0.00	0.00	NA	0.00	94.70	0.03
		Model 6	Ignoring clustering	-2.1(1.5)	0.00	0.00	0.00	NA	0.00	94.60	0.03
		Model 8	One-step including indicator variable.	-2.1(1.5)	0.00	0.00	0.00	NA	0.00	94.50	0.03

Appendix E

Appendix E1: STATA code for possible FP IPD meta-analysis models

Two-step model

FP1 IPD meta-analysis of logistic regression model

fracgen age -2 -1 -0.5 0 0.5 1 2 3

logit dead age10_1
logit dead age10_2
logit dead age10_3
logit dead age10_4
logit dead age10_5
logit dead age10_6
logit dead age10_7
logit dead age10_8

Then I selected the best model that has the highest sum Chi-square statistics for the 4 IPD meta-analysis and I applied the second step of meta-analysis by using fixed effect methods to estimate the pooled effect size of the PF(age), gives as

.metan beta s.e. (beta)

One-step model

\\Accounting for the clustering of patients within-studies by using indicator variables for the studies with Fractional polynomial model with power=1\\

fracpoly logit dead age10 f11 f12 f13 f14 , nocons degree(1) log center(no) noscaling compare
\\Accounting for the clustering of patients within-studies by using indicator variables for the studies with Fractional polynomial model with power=2\\

fracpoly logit dead f11 f12 f13 f14 age10, nocons degree(2) compare
fracplot age, msize(small)

Appendix F

Appendix F1: STATA code for possible FP IPD meta-analysis models

Contour funnel plot

```
confunnel coef se, legend(pos(6))
```

Cumulative meta-analysis forest plot

```
metacum coef se,lcols(samplesize cumulative standard error)
```

Egger's test: Small –study effect

```
metabias coef  lose, egger
or,
regress coef se, [aweight=1/var]
```

Orginal Egger test:

```
regress  coef var[aweight=1/var]
```

List of References

1. Chap TL. Introductory Biostatistics.: Wiely, 2003.
2. Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. . *BMJ* 2009;339:b4184 doi: 10.1136/bmj.b4184.
3. Altman D. Systematic reviews of evaluations of prognostic variables. *BMJ* 2001;doi: 10.1136/bmj.323.7306.224:224-8.
4. Sauerbrei W. Prognostic factors: confusion caused by bad quality design, analysis and reporting of many studies Current Research in Head and Neck Cancer. Advances in Oto-Rhino-Laryngology. *Karger: Basel* 2005; 62:184–200.Bier H (ed).
5. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer.* 1994; 69(6):979–85.
6. Holländer N, Sauerbrei W. . On Statistical Approaches for the Multivariable Analysis of Prognostic Marker Studies. *Statistics for Industry and Technology.* 2007;DOI: 10.1007/978-0-8176-4542-7_2:19-38.
7. Kyzas P, Loizou KT, Ioannidis J P A. Selective Reporting Biases in Cancer Prognostic Factor Studies. *JNCI J Natl Cancer Inst* 2005;97 (14) doi: 10.1093/jnci/dji184,1043-55.

8. Altman D, Lausen B, Sauerbrei W, Schumacher M. Dangers of Using "Optimal" Cutpoints in the Evaluation of Prognostic Factors. *Journal of the National Cancer Institute* 1994;86(11).
9. Tierney J, Clarke M, Stewart LA. Is there bias in the publication of individual patient data meta-analyses? *Int J Technol Assess Health Care.* 2000;16(2):657-67.
10. Simmonds M, Higgins JPT,Tierneyb JF., et al. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *clinical trial* 2005;2(3)doi: 10.1191/1740774505cn087oa 209-17.
11. Broeze KA, Opmeer BC ,Fulco VDV,et al. Individual patient data meta-analysis: a promising approach for evidence synthesis in reproductive medicine.*Human Reproduction Update* 2010;16(6) doi:10.1093/humupd/dmq043:561–67.

12. Rachet B, Maringe C, Nur U, et al. Population-based cancer survival trends in England and Wales up to 2007: an assessment of the NHS cancer plan for England. *Lancet Oncol* 2009 10(4):351-69.
13. Sauerbreia W, Hollnder 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 -Theory and Methods.* 2006;35(7):1333-42.
14. Look MP, Van Putten W L,Duffy M J,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(2):116-28.
15. Altman D, Riley R D. Primer: An evidence-based approach to prognostic markers. *Nature Clinical Practice Oncology.* 2005(2(9)):466-72.
16. Braun S, Vogl FD,Naume B, et al. A pooled analysis of bone marrow micrometastasis in breast cancer. *New England Journal of Medicine.* 2005(353(8)):793-802.
17. Galea M, Blamey RW,Elston CE, Ellis IO. he Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Research and Treatment* 1992;22:207-19.
18. Parmar M, Torri V, Stewart L. Extracting summary statistics to perform metaanalyses of the published literature for survival endpoints. . *Statistics in Medicine.* 2008;100(2):92-97.
19. Riley R, Sauerbrei W,Altman DG. Prognostic markers in cancer: the evolution of evidence from single studies to meta analysis. *British Journal of Cancer.* 2009;100(8):1219–29.
20. Collaboration C. Cochrane Handbook for Systematic Reviews of Interventions. In: Higgins JPT, Green, S., editor, 2011.

21. Compton C, Fielding LP, Burgart LJ, et al., . Prognostic Factors in Colorectal Cancer. . *Arch Pathol Lab Med* 2000;124(7):979-94.
22. Riley R, Abrams KR, Sutton AJ, et al, . Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. *Br J Cancer* 2003;88:1191–98
23. McShane L, Altman, DG, et al. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK). *Journal of the National Cancer Institute.* 2005;97(16):416-22.
24. Altman D, De Stavola, BL, Love, SB, Stepniewska, KA. Review of survival analyses published in cancer journals. *Br J Cancer.* 1995;72(2):511-8.
25. Altman D, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Resarch and Treatment.* 1998;52(1-3):289-303.
26. Hayden J, Cote P,Steenstra I A,Bonbardier C. Identifying phases of investigation helps planning, apprassing, and applying the results ofexplanatory prognosis studies. *Journal of Clinical Epidemology* 2008;61(6):552-60.
27. Buettner P, Garbe C, Guggenmoos-Holzmann I. Problems in defining cutoff points of continuous prognostic factors: Example of tumor thickness in primary cutaneous melanoma. *Journal of Clinical Epidemiology* 1997;50(11):1201-10.
28. Altman D, Patrick R. The cost of dichotomising continuous variables. . *BMJ* 2006;332 : 1080 doi: 10.1136/bmj.332.7549.1080
29. Lausena B, Schumacherb M. Evaluating the effect of optimized cutoff values in the assessment of prognostic factors *Computational Statistics & Data Analysis* 1999;21(3):307-26

30. Hukkelhoven CW, Steyerberg EW, Rampen A J,et al. Patient age and outcome following severe traumatic brain injury: an analysis of 5600 patients. *Journal of Neurosurgery* 2003;99(4):666-73.
31. Steyerberg EW, Mushkudiani N, Perel P, et al. Predicting outcome after traumatic brain injury: Development and international validation of prognostic scores based on admission characteristics. *PLoS Medicine.* 2008(5(8)):1251-61.
32. Kleinbaum D. Logistic Regression- A self learning text: springer, 1994.
33. McCullagh P, Nelder. JA. Generalized Linear Models: Chapman and hall, 1989.
34. Lee E, Wang J. Statistical Methods for Survival Data Analysis: John Wiley, 2003.
35. Hosmer DW, Lemeshow S., editor. *Applied survival analysis: regression modeling of time to event data,* 1999.

36. Selvin S, editor. *Survival Analysis for Epidemiologic and Medical Research*: Cambridge University Press, 2008.
37. Collet D, editor. *Modelling Survival Data in Medical Research*: Chapman and Hall, 1994.
38. Spruance SL, Reid J E, Grace M, Samore M. Hazard Ratio in Clinical Trials. *American Society for Microbiology* 2004;48(DOI: 10.1128/AAC.48.8.):2787–92.
39. Altman DG. Practical statistics for medical research: Chapman and hall, London., 1991.
40. Kleinbaum DG, Klein M., editor. *Survival Analysis: A Self-Learning Text.*: Springer Science and Business Media, Inc., 2005.
41. Schoenfeld D. Partial residuals for the proportional hazards regression-model. *Biometrika* 1982;69:239–41
42. Grambsch PM, Therneau T.M. Proportional hazards tests and diagnostics basedon weighted residuals. *Biometrika* 1994;81:515-26.
43. Hosmer D.W. , S. LJ, editors. *Applied Survival Analyisis*: Wiely series in probability and statistics,, 1999.

44. Falagas ME, Manta K G, Ntziora F, Vardakas K Z. Linezolid for the treatment of patients with endocarditis: A systematic review of the published evidence. *Journal of Antimicrobial Chemotherapy.* 2006(58(2)):273-80.

45. Egger M, Smith, GD, et al, editor. *Systematic Reviews in Health Care: Meta-Analysis in Context*, 2001.
46. Borenstein MHL, Higgins J P, Rothstein HR., editor. *Introduction to Meta-Analysis*, 2009.
47. Higgins J, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine.* 2002;21(11):1539–58.
48. Sterne ACJ, Bradburn M J, Egger M. Meta-analysis in Stata. *Systematic Reviews* 2011;18-cpp:347.
49. Harris R, Bradburn MJ, Deeks J J. metan: fixed- and random-effects meta-analysis. *The Stata Journal* 2008;8(1):3–28.
50. Riley R, Heney D, Jones DR, et al. A Systematic Review of Molecular and Biological Tumor Markers in Neuroblastoma. *Clin Cancer Res* 2004;DOI:10.1158/1078-0432.CCR-1051-2:4-12.
51. Altman D.G. LGH. Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Research and Treatment,* 2004;52:289-303.
52. Kyzas PA, Denaxa-Kyza D, Ioannidis JPA. Quality of Reporting of Cancer Prognostic Marker Studies: Association With Reported Prognostic Effect. *J Natl Cancer Inst* 2007;99 236 – 43.
53. Riley R, Abrams K, Lambert P, Sutton A, Altman D. Where Next for Evidence Synthesis of Prognostic Marker Studies? Improving the Quality and Reporting of Primary Studies to Facilitate Clinically Relevant Evidence-Based Results *Statistics for Industry and Technology* 2007; DOI: 10.1007/978-0-8176-4542-7_3 39-58.
54. Riley R, Simmonds MC,Look M P. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. *Journal of Clinical Epidemiology* 2007;60(5):431-39.
55. Riley R. Commentary: Like it and lump it? Meta-analysis using individual participant data. *International Journal of Epidemiology.* 2010;1–3 doi:10.1093/ije/dyq129:1359-61.
56. Cooper H, Patall EA. The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods* 2009;14(2):165-76.
57. Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate data. *BMC Medical Research Methodology.5, 2005.Article Number: 14.Date of Publication: 25 Apr 2005.* 2005:-.
58. Smith CT, Williamson PR, Marson AG. An overview of methods and empirical comparison of aggregate data and individual patient data results for investigating heterogeneity in meta-analysis of time-to-event outcomes. *Journal of Evaluation in Clinical Practice.11(5)()(pp 468-478), 2005.Date of Publication: Oct 2005.* 2005(5):468-78.
59. Tudur SC, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics in Medicine.24(9)()(pp 1307-1319), 2005.Date of Publication: 15 May 2005.* 2005(9):1307-19.
60. Riley R, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010; 340:c221 doi: 10.1136/bmj.c221.
61. Sud S, Douketis J. The devil is in the details…or not? A primer on individual patient data meta-analysis. *Evid Based Med* 2009;14(100-101 doi:10.1136/ebm.14.4.100).
62. Oxman A, Clarke MJ, Stewart LA. Meta-analyses Using Individual Patient Data Are Needed. *The Journal of American Statistical Association* 1995;274(10) doi: 10.1001/jama.1995.03530100085040:845-46.
63. Morton N, Keating JL. Different Meta-Analysis Methods, Different Answers: The Case of Exercise for Older Acute Medical Patients. *The Open Geriatric Medicine Journal* 2009;2:9-11.
64. Stewart LA, Tierney, J.F. To IPD or not to IPD? : Advantages and Disadvantages of Systematic Reviews Using Individual Patient Data. *Evaluation and The Health Professions* 2002;25(76)(DOI: 10.1177/016327870202500l006):76-97.
65. Ioannidis J, Rosenberg PS, Goedert JJ, O'BrienTR. Commentary: meta-analysis of individual participants' data in genetic epidemiology. *Am J Epidemiol* 2002;156(3):204-10.
66. Turner RM OR, Yang M, Goldstein H,Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Stat Med,* 2000;19(24):3417–32.

67. Higgins JP, Whitehead A, Turner RM, Omar RZ, SG T. Meta-analysis of continuous outcome data from individual patients. *Stat Med* 2001;20(15)(2219–2241).
68. Whitehead A, Omar RZ, Higgins JP, et al. Meta-analysis of ordinal outcomes using individual patient data. *Stat Med* 2001;20(15):2243-60.

69. Riley R, Lambert PC, Staessen JA, et al, . Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine.* 2008(27(11)):1870-93.
70. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *Jama* 2000;283(15):2008-12.
71. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet* 1999;354(9193):1896-900.
72. Trivella M, Pezzella F, Pastorino U,Harris AL,Altman DG. Microvessel density as a prognostic factor in non-small-cell lung carcinoma: a meta-analysis of individual patient data. *Lancet Oncology.* 2007(8(6)):488-99.
73. MeRGE1. Independence of restrictive filling pattern and LV ejection fraction with mortality in heart failure: An individual patient meta-analysis. *Eur J Heart Fail* 2008;10 (8):786-92.
74. Murray G, Butcher I, McHugh GS, Lu J, Mushkudiani N A, Maas AIR, et al. Multivariable prognostic analysis in traumatic brain injury: Results from the IMPACT study. . *Journal of Neurotrauma* 2007;24(2): 329-37.
75. Butcher I, Maas, AIR, Lu J, Marmarou A, Murray GD, Mushkudiani NA, et al. . Prognostic value of admission blood pressure in traumatic brain injury: Results from the IMPACT study. *Journal of Neurotrauma* 2007;24(2):294-302.
76. Beek V, Mushkudiani NA, Steyerberg E W, et al., . Prognostic value of admission laboratory parameters in traumatic brain injury: Results from the IMPACT study. *Journal of Neurotrauma* 2007;.24(2)(2):315-28.
77. Maas A, Steyerberg EW, Butcher I, Dammers R, Lu J, Marmarou A. Prognostic value of computerized tomography scan characteristics in traumatic brain injury: Results from the IMPACT study. *Journal of Neurotrauma* 2007;24(2):303-14.
78. Mushkudiani N, Engel DC, Steyerberg EW, Butcher I, Lu J, Marmarou A, et al. Prognostic value of demographic characteristics in traumatic brain injury: Results from the IMPACT study. *Journal of Neurotrauma* 2007;24(2):259-69.
79. McHugh G, Engel DC, Butcher I, et al. , . Prognostic value of secondary insults in traumatic brain injury: Results from the IMPACT study. *Journal of Neurotrauma.* 2007;24(2)(2):287-93.
80. Yap Y, DuongT,Bland M, et al., . Potential demographic and baselines variables for risk stratification of high-risk post-myocardial infarction patients in the era of implantable cardioverter-defibrillator - A prognostic indicator. *International Journal of Cardiology.* 2008;126(1):101-07.
81. Thakkinstian A, Dmitrienko S,Gerbase-DeLima M, et al. , . Association between cytokine gene polymorphisms and outcomes in renal transplantation: A meta-analysis of individual patient data. *Nephrology Dialysis Transplantation.* 2008;23(9)(9):3017-23.
82. Koopman L, Hoes A W, Glasziou PP, et al., . Antibiotic therapy to prevent the development of asymptomatic middle ear effusion in children with acute otitis media: A meta-analysis of individual patient data. *Archives of Otolaryngology - Head and Neck Surgery.* 2008(2):128-32.
83. Lanterna L, Ruigrok Y,Alexander S, et al. Meta-analysis of APOE genotype and subarachnoid hemorrhage: Clinical outcome and delayed ischemia. *Neurology* 2007;69(8)(8):766-75.
84. MeRGE2. Independent prognostic importance of a restrictive left ventricular filling pattern after myocardial infarction: an individual patient meta-analysis: Meta-Analysis Research Group in Echocardiography acute myocardial infarction.[see comment]. *Circulation* 2008;117(20):2591-98.

85. Downing M, Lau F, Lesperance M, et al., . Meta-analysis of survival prediction with Palliative Performance Scale. *Journal of Palliative Care.* 2007;23(4):245-52.
86. Noordzij J, Lee SL,Bernet VJ, et al. Early Prediction of Hypocalcemia after Thyroidectomy using Parathyroid Hormone: An Analysis of Pooled Individual Patient Data from Nine Observational Studies. *Journal of the American College of Surgeons.* 2007;205(6):748-54.
87. Rovers MM, Glasziou P, Appelman C L, et al., . 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.119(3)* 2007;119(3):579-85.
88. Schaich M, Schlenk R F, Al-Ali HK, et al. Prognosis of acute myeloid leukemia patients up to 60 years of age exhibiting trisomy 8 within a non-complex karyotype: Individual patient data-based meta-analysis of the German Acute Myeloid Leukemia Intergroup. *Haematologica.92* 2007;92(6):763-70.
89. Goetz C, Wuu J,McDermott M P, et al. Placebo response in Parkinson's disease: Comparisons among 11 trials covering medical and surgical interventions. *Movement Disorders* 2008;23(5)(5):690-99.
90. Warkentin TE, Sheppard J A,Sigouin C S, et al. Gender imbalance and risk factor interactions in heparin-induced thrombocytopenia. *Blood.108(9)* 2006;108(9):2937-41.
91. Sylaja PN, Dong W, Grotta JC, et al. Safety outcomes of Alteplase among acute ischemic stroke patients with special characteristics. *Neurocritical Care.* 2007;6(3):181-85.
92. MeRGE1. Independent prognostic importance of a restrictive left ventricular filling pattern after myocardial infarction: an individual patient meta-analysis: Meta-Analysis Research Group in Echocardiography acute myocardial infarction.[see comment]. *Circulation* 2008;117(20):2591-98.
93. MeRGE2. Independence of restrictive filling pattern and LV ejection fraction with mortality in heart failure: An individual patient meta-analysis. *Eur J Heart Fail* 2008;10 (8):786-92.
94. Higgins J, Whitehead A,Turner RM,et al. Meta-analysis of continuous outcome data from individual patients. *Stat Med.* 2001;20(15)(2219–2241).
95. Higgins JP, Thompson, S.G., Deeks, J.J., Altman, D.G. Measuring inconsistency in meta-analyses. *BMJ* 2003;327(7414):557-60.
96. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7(3):177-88.
97. Rovers MM, Glasziou P, Appelman CL, Burke P, McCormick DP, Damoiseaux RA, et al. 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.119(3)()(pp 579-585), 2007.Date of Publication: Mar 2007.* 2007(3):579-85.
98. Look M, Foekens J A. Clinical relevance of the urokinase plasminogen activator system in breast cancer. *APMIS* 1999;107:150-09.
99. Jones PA, Riley RD, Williamson PR, Whitehead A. Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. *Clinical trial* 2009;6(1)(doi: 10.1177/1740774508100984).
100. Sutton A J AK, Jones DR, Sheldon TA, Song F., editor. *Methods for Meta-analysis in Medical Research*: John Wiley, 2000.
101. Thompson SG, Kaptoge S, White I, Wood A, Perry P, Danesh J. . Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. *International Journal of Epidemiology* 2010;39(5):1345-59
102. Simmonds MC, Higgins J PT. Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. *Statistics in Medicine.* 2007(26(15)):2982-99.
103. Tudur-Smith C JA, Alfirevic Z, Williamson PR. Individual patient data meta-analysis : Cervical stitch (cerclage) for preventing pregnancy loss in women. *BMC Pregnancy and Childbirth* 2005;5 (5) doi:10.1186/1471-2393-5-5.

104. Rothstein HR, Sutton AJ, Borenstein M., editor. *Publication Bias in Meta-Analysis Preventation, Assessment and Adjustements*: John Wiley& Sons,Ltd, 2005.
105. Terrin N, Schmid C H, Lau J, Olkin I. , . Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine* 2005;22:2113-26.
106. Egger M, Smith GD,Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629-34.
107. Stewart L, Parmar MK. Meta-analysis of the literature or of individual patient data: is there a difference? *The Lancet* 1993;341(8842):418–22.
108. Berlin JSJ, Schmid CH. Individualpatient- versus group-leveldata meta-regressions for the investigation of treatment e*ect modi,ers: ecological bias rears its ugly head. *Statistics in Medicine* 2002;21(DOI: 10.1002/sim.1023):371–87.
109. Williamson P, Smith CT, Hutton JL, Marson AG. . Aggregate data meta-analysis with time-to-event outcomes. *Statistics in Medicine* 2002;21(DOI: 10.1002/sim.1303):3337–51.
110. Steyerberg E, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infraction: should we adjust for baseline characteristics? *Am Heart J* 2000;139(5):745-51.
111. Steyerberg E, editor. *Clinical Prediction Models*: Springer Science, 2010.
112. Riley R, Steyerberg EW. Meta-analysis of a binary outcome using individual participant data and aggregate data. *Research Synthesis Methods* 2009;DOI:10.1002/JRSM.4.
113. Borenstein M, Hedges LV, Higgins JPT,Rothstein HR., editor. *Introduction to Meta-Analysis*: John Wiley & Sons, Ltd, 2009.
114. Leandro G, Gallus, G., editor. *Meta-analysis in medical research*: Blackwell Publishing Ltd, 2005.
115. Hartung J, Knapp G, Sinha BK. Statistical Meta-Analysis with Applications: John Wiley & Sons, 2008.
116. Whitehead A, editor. *Meta-Analysis of Controlled Clinical Trials*: John Wiley & Sons Ltd, 2002.
117. Kulinskaya E, Morgenthaler S, Staudte RG, editor. *Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence*: John Wiley & Sons Ltd,, 2008.
118. Chu H, Guo H, Zhou Y. Bivariate Random Effects Meta-Analysis of Diagnostic Studies Using Generalized Linear Mixed Models. *Medical in Decision Making* 2010;DOI: 10.1177/0272989X09353452.
119. Ishak K, Platt RW, Joseph L,Hanley AJ. Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. *Statistics in Medicine* 2007.
120. Jackson D, Riley R, White I R. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine* 2011;DOI: 10.1002/sim.4172
121. Riley R, Abrams K R , Lambert P C, Sutton A J. An evaluation of bivariate random-effect meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* 2007;26(DOI:10.1002/SIM.2524):78-97.
122. Riley R, Abrams KR, Sutton A J , Lambert P L, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *Medical Research Methodology* 2007;7:3doi:10.1186/1471-2288-7-3
123. R.D. R. Multivariate meta-analysis: the effect of ignoring wuthin-study correlation. *Royal Statistical Society* 2009;172(Series A):789-811.
124. Thompson SG TR, Warn DE. Multilevel models for meta-analysis, and their application to absolute risk differences. *Statistical Methods in Medical Research* 2001;10:375–92.
125. Littell R, Pendergast J, Natarajan R. Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine* 2000;19:1793-819.
126. Gelman A, Hill J., editor. *Data Analysis Using Regression and Multilevel /Hierarchical Models*: Cambridge University Press, 2006.
127. Harbord R. M. WP. metandi: Meta–analysis of diagnostic accuracy using hierarchical logistic regression. *The Stata Journal* 2009;9(2):211-29.

128. StataCorp L, editor. *STATA Longitudinal Data/ Panel Data Reference Manual Release 11* A Stata Press, 2009.
129. Goldstein H, Yang M, Omar RZ, Turner RM, Thompson SG. . Meta-analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society, Series C—Applied Statistics* 2000;49:399--412.
130. Mathew T, Nordström K. Comparison of One-Step and Two-Step Meta-Analysis Models Using Individual Patient Data. . *Biometrical Journal* 2010;52(2):271–87 DOI: 10.1002/bimj.200900143.
131. Hogan J, Blazar A S. . Hierarchical logistic regression models for clustered binary outcomes in studies of IVF-ET. *Fertility and Sterility,* ;73(3):575-81.

132. Burton A, Altman DG, Royston P, Holder R L. The design of simulation studies in medical statistics. *Statistics in Medicine.* 2006;25, DOI: 10.1002/sim.2673:4279–92.
133. Johnson R. Assessment of Bias with Emphasis on Method Comparison. *The Clinical Biochemist Reviews* 2008;29.
134. Muthén L. How to use a Monte Carlo study to decide on sample size and determine power. . *Structure Eguation Modelling* 2002;4:599-620.
135. Lehmann EL, Casella G. Theory of Point Estimation. New York: Springer, 1998.
136. Altman DG DJ. Meta-analysis,Simpson's paradox, and the numver needed to treat. *BMC Medical Research Methodology* 2002;2(3).
137. Peters T, Richards SH, Bankhead CR, Ades AE, Sterne JA. Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. *Int J Epidemiol.* 2003;32(5):840-6.
138. Bland M. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Medical Research Methodology* 2004;4, 21.
139. Clarke P. When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *J Epidemiol Community Health*. 2007;62:752-58.
140. Hedeker D, Gibbons RD, Flay BR. Random-Effects Regression Models for Clustered Data With an Example from Smoking Prevention Research. *Journal of Consulting and Clinical Psychology* 1994;62(4):757-65.
141. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials.*Prog Cardiovasc Dis* 1985;27(5):335-71.

142. Mantel N, Hanensizel, w., MANTEL N HW. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst.* 1959;22(4): 719-48.
143. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. . *Applied Statistics* 1994;43:429- 67.
144. Long J, Ryoo J. Using fractional polynomials to model non-linear trends in longitudinal data. *British Journal of Mathematical and Statistical Psychology.*2010;63 (1):177-203.

145. Sauerbrei W, Royston, P. . Determination of Functional Relationships for Continuous Variables by Using a Multivariable Fractional Polynomial Approach. *Lecture Notes in Computer Science* 2002;DOI: 10.1007/3-540-36104-9_6 53-60.
146. Royston P, Sauerbrei,W. . Improving the robustness of fractional polynomial models by preliminary covariate transformation: A pragmatic approach. . *Computational Statistics & Data Analysis* 2007;51:4240 – 53.
147. Royston P, Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Statistics in Medicine* 2003;22(4):639-59.
148. Royston P, Sauerbrei W., editor. *Multivariable Model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*: John wiley & Sons Ltd, 2008.

149. Amber G, Royston P Fractional polynomial modelselection procedures: investigation of Type I error rate. *Journal of Statistical Computation and Simulation* 2001;69:89-108.
150. Saurerbrei W, and Rayston 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.
151. Rota M, Bellocco R, Scotti L,et al. Random-effects meta-regression models for studying nonlinear dose–response relationship, with an application to alcohol and esophageal squamous cell carcinoma. *Statistics in Medicine.* 2010;DOI: 10.1002/sim.4041:2679–87.
152. Fedirko V, Tramacere I, BagnardiV, et al. Alcohol drinking and colorectal cancer risk: an overall and dose–response meta-analysis of published studies. *Annals of Oncology* 2011;doi:10.1093/annonc/mdq653.
153. Bagnardi V ZA, Quatto P, Corrao G. Flexible Meta-Regression Functions for Modeling Aggregate Dose-Response Data, with an Application to Alcohol and Mortality. *American Journal of Epidemiology* 2004;159 (11)(DOI: 10.1093/aje/kwh142).
154. Ahmed I. Assessment of Publication and Availability Bias in IPD meta-analysis of randomised trials. *MSc Medical statistics*. University of Leicester;Departement of Health Sciences, 2010.
155. Newcombe RG. Towards a reduction in publication bias. *BMJ* 1987;295(6599):656–59.
156. Angell M. Negative studies. *N Engl J Med* 1989 321(7):464-6.
157. Riley RD. Commentary: Like it and lump it? Meta-analysis using individual participant data. *International Journal of Epidemiology 2010;1–3* 2010;doi:10.1093/ije/dyq129.
158. Sterne J, Egger M. . Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of Clinical Epidemiology* 2001;54(doi:10.1016/S0895-4356(01)00377-8):1046–55.
159. Duval S, Tweedie R. Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *Biometrics* 2000;56:455-63.
160. Sterne J, Sutton, AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;342:d4002 doi: 10.1136/bmj.d4002.
161. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Statistics In Medicine* 2008;20:641-54.
162. Schulz K, Chalmers I, Hayes RJ,Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273(5):408-12.
163. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *J Clin Epidemiol* 2005;58:894-901.
164. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology* 2008;61(991-996).
165. Moreno SG, Sutton,A J, Ades,AE,Stanley, T D. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology* 2009;9:2 doi:10.1186/1471-2288-9-2.
166. Peters JL, Sutton A J, Jones D R , Abrams KR, Rushton L. . Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine.* 2007;26 DOI: 10.1002/sim.2889:4544–62.
167. Duval S, Tweedie R. A Nonparametric "Trim and Fill" Method of Accounting for Publication Bias in Meta-Analysis. *Journal of the American Statistical Association* 2010;95 89-98.
168. Palma S, Delgado-Rodriguez M. Assessment of publication bias in meta-analyses of cardiovascular diseases. *J Epidemiol Community Health.* 2005;59(10):864-9.
169. Sutton A, Duval S J, Tweedie R L, Abrams K R, Jones D R. Empirical assessment of effect of publication bias on meta analyses. *BMJ* 2000;320.

170. Kyzas PA, Denaxa-Kyza D, JP I. Almost all articles on cancer prognostic markers report statistically significant results. *Eur J Cancer* 2007a;43:2559–79
171. Hutchon D. Publishing raw data and real time statistical analysis on e-journals. . *BMJ* 2001;530 doi: 10.1136/bmj.322.7285.530.
172. Moher D, Liberati A, Tetzlaff J, Altman DG. (for the PRISMA Group). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.
173. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006;25(1):127-41.
174. Sauerbrei W. Prognostic Factors - Confusion caused by bad quality of design, analysis and reporting of many studies. In: Bier H, editor. *Current Research in Head and Neck CancerAdvances in Oto-Rhino-Laryngology*. Basel: Karger, 2005:184-200.
175. Koopman L, van der Heijden GJ, Grobbee DE, Rovers MM. Comparison of methods of handling missing data in individual patient data meta-analyses: an empirical example on antibiotics in children with acute otitis media. *American journal of epidemiology*;167(5):540-5.
176. Higgins J, White IR, Wood AM. Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clinical Trials*;5 225–39.
177. Turner R, Omar RZ, Thompson SG. Modelling multivariate outcomes in hierarchical data, with application to cluster randomised trials. *Biometrical Journal* 2006;48(3):333-45.