

A guide to systematic review and meta-analysis of prognostic factor studies

Richard D Riley^{1#*}, Karel GM Moons^{2,4#}, Kym IE Snell¹, Joie Ensor¹,
Lotty Hooft^{2,4}, Douglas G Altman³, Jill Hayden⁴, Gary S Collins³, Thomas
PA Debray^{2,5}

Contact details:

* corresponding author: Professor of Biostatistics; e-mail: r.riley@keele.ac.uk;

Tel: +44 (0) 1782 733905 Fax: +44 (0) 1782 734719

both authors contributed equally

¹ Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, UK. ST5 5BG

² Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

³ Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK. OX3 7LD.

⁴ Centre for Clinical Research, 5790 University Ave, Halifax, Nova Scotia, Canada. B3H 1V7

⁵ Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

WORD COUNT: 6000

Summary Points

- Primary studies to identify prognostic factors are abundant, but often have conflicting findings and variable quality. This motivates systematic reviews and meta-analyses to identify, evaluate and summarise prognostic factor studies and their findings.
- A clear review question should be framed using a PICOTS system, and a transparent search undertaken for eligible articles. Broad search strings may be required, leading to a large number of articles to screen.
- A data extraction phase is needed to obtain the relevant information from each study. A modification of the CHARMS checklist can be used (CHARMS-PF).
- The QUIPS tool can be used to examine each study's risk of bias. Unfortunately many primary studies will have a high risk of bias due to poor standards of design, conduct, analysis and reporting. Applicability of a study should also be checked.
- If appropriate, meta-analysis can be used to combine prognostic effect estimates (such as hazard ratios or odds ratios) across studies, to produce an overall summary of a factor's prognostic effect. Between-study heterogeneity should be expected and accounted for.
- Ideally separate meta-analyses should be performed for unadjusted and adjusted results; the latter is more important, to examine a factor's independent prognostic value over and above (i.e. after adjustment for) existing prognostic factors.
- Separate meta-analyses may also be required for each method of measurement (for factors and outcomes), each approach to handling continuous factors, and each type of estimate (e.g. hazard ratios, odds ratios).
- Publication bias is a major threat to the validity of meta-analyses based on published evidence, and may cause small-study effects (asymmetry on a funnel plot).
- REMARK and PRISMA can be used to guide the reporting of the systematic review and meta-analysis, and the degree of confidence in the summary results from the review may be examined using adapted forms of GRADE for interventions and diagnostic test accuracy studies.
- Availability of individual participant data (IPD) may alleviate some of the challenges.

Standfirst

Prognostic factors are associated with the risk of future health outcomes in individuals with a particular health condition or clinical start-point. Research to identify genuine prognostic factors is important, as they may help improve risk stratification, treatment decisions, and the design of randomised trials. However, thousands of prognostic factor studies are published each year, often with variable quality and inconsistent findings. This motivates the need for systematic reviews and meta-analyses that summarise the evidence about the prognostic value of particular factors. Here, we describe the key steps involved in this review process.

Introduction

Systematic reviews and meta-analyses are common in the medical literature, routinely appearing in specialist and general medical journals, and forming the cornerstone of Cochrane. The majority of systematic reviews focus on summarising the benefit of one or more therapeutic interventions for a particular condition. However, they are also important for summarising other evidence, such as summarising the accuracy of screening and diagnostic tests,¹ the aetiological association of risk factors for disease onset, and the prognostic ability of bespoke factors and (bio)markers. The latter arise from prognosis studies, which aim to examine and predict future outcomes (such as *death, disease progression, or medical complications such as pre-eclampsia*) in those with(in) a particular health condition or start-point (such as receiving a certain *diagnosis, undergoing surgery, or being pregnant*).

The PROGRESS (PROGnosis RESearch Strategy) framework defines four types of prognosis research objectives: (i) to summarise overall prognosis (e.g. overall risk or rate) of health outcomes for groups defined by a particular health condition,² (ii) to identify prognostic factors associated with changes in health outcomes,³ (iii) to develop, validate and examine the impact of prognostic models for individualised prediction of such outcomes,⁴ and (iv) to identify predictors of an individual's response to treatment.⁵ Each topic area requires specific methods and tools for conducting a systematic review and meta-analysis. Two recent articles provided a guide to undertaking reviews and meta-analysis of prognostic (prediction) models.^{6,7} Here, we focus on prognostic factors, which is the most common type of prognosis research.

A prognostic factor is any variable that, among people with a given health condition, is associated with the risk of a subsequent health outcome. Different values or categories of a prognostic factor are associated with a better or worse prognosis, i.e. of future health outcomes. For example, in many cancers tumour grade at the time of histological diagnosis is a prognostic factor because it is associated with time to subsequent disease recurrence or death. Each grade represents a group of patients with a different average prognosis, and the risk or rate (hazard) of the outcome increases with higher grades. Many routinely collected patient characteristics are prognostic, such as sex, age, body mass index, smoking status, blood pressure, co-morbidities and symptoms. Many researched prognostic factors are biomarkers, which include a diverse range of blood, urine, imaging, electrophysiological, and physiological variables.

Prognostic factors have many potential uses, including aiding treatment decisions, improving individual risk prediction, providing novel targets for new treatment, and enhancing the design and

analysis of randomised trials.³ This motivates so-called ‘prognostic factor research’ to identify genuine prognostic factors (sometimes also called ‘predictor finding studies’⁸).⁹ Thousands of such studies are published each year, but they often have variable quality and inconsistent findings. This motivates the need for systematic reviews and meta-analyses that summarise the evidence about the prognostic value of particular factors.¹⁰⁻¹² In this article, we provide a step-by-step guide to conducting such reviews. Our aim is to help researchers understand the key principles, methods, and challenges of conducting reviews of prognostic factor studies, to produce robust evidence-based summaries about prognostic factors.

Step 1: Defining the research question

The first step is to define the review question. Reviews of prognostic factor studies fall within the remit of type 2 of the PROGRESS framework,² as they aim to summarise the prognostic value of a particular factor (or each of multiple factors) within a particular disease field for relevant health outcomes and time-points. Some reviews may be broad. For example, Riley et al. aim to identify *any* prognostic factor for overall and disease-free survival in children with neuroblastoma or Ewing’s sarcoma.¹³ Other reviews may have a narrower focus. For example, Hemingway et al. aim to summarise the evidence for whether C-reactive protein (CRP) is a prognostic factor for fatal and nonfatal events among patients with stable coronary disease.¹⁴ This CRP review will be used as a running illustrative example throughout this article.

CHARMS (**C**hecklist for critical **A**ppraisal and data extraction for systematic **R**reviews of prediction **M**odelling **S**tudies) provides guidance for formulating a review question and a checklist for extracting data and critically appraising included studies.¹⁵ Though developed,¹⁵ and further refined,⁶ for reviews of prediction model studies, it can also be used to define and frame the question for reviews of prognostic factor studies. CHARMS¹⁵ and subsequent improvements⁶ propose a modification (called PICOTS) of the traditional PICO system (Population, Index intervention, Comparison, and Outcome) used in systematic reviews of therapeutic intervention studies, by additionally considering Timing and Setting (see Figure 1). In the context of prognostic factor reviews, the P of Population and O of Outcome remain largely the same, but now the I refers to Index prognostic factor(s) and the C refers to other prognostic factors that can be considered as Comparators in some way(s). For example, the aim may be to compare the prognostic ability of a certain index factor to one or more other (i.e. comparator) prognostic factors; or to investigate the adjusted prognostic value of a particular index factor, i.e. over and above (adjusted for) other (i.e.

comparator) prognostic factors. If the only aim is to summarise the unadjusted prognostic effect of a particular index factor, which we do not generally recommend, then there is actually no comparator factor being addressed. The T denotes Timing and actually refers to two concepts of time: (i) at what time-point the prognostic factors under review are to be used (i.e. the time point of at which prognosis information is required) and (ii) over what time period the outcome(s) are predicted by these factors. The S of Setting refers to the setting or context in which the index prognostic factor(s) are to be used as the prognostic ability of a factor may change across healthcare settings."

An important component of reviews addressing prognostic factors is whether unadjusted or adjusted estimates of the index prognostic factor(s) will be summarised, or both. We recommend that reviewers primarily focus on *adjusted* prognostic effects, as these reveal whether a certain index factor contributes independently to the prediction of the outcome, irrespective of (i.e. adjusted for) other prognostic factors. In particular, usually for each clinical scenario there are so-called 'established' or 'conventional' prognostic factors that are always measured. Therefore, for prognostic factors under review, it is important to understand whether they contribute additional (sometimes called 'independent') prognostic information to these routinely measured ones. This means that adjusted (and not unadjusted or crude) prognostic effect estimates need to be estimated and reported in primary prognostic factor studies. Such independent effects are typically derived from a multivariable regression model containing both the established prognostic factors plus each index prognostic factor of interest. For example, consider a logistic regression of a binary outcome including three adjustment factors (A_1 , A_2 and A_3 , say) and one new index prognostic factor (X_1 , say), which is expressed as:

$$\ln(p/(1 - p)) = \alpha + \beta_1 A_1 + \beta_2 A_2 + \beta_3 A_3 + \beta_4 X_1$$

Here, p is the probability of the outcome, and after estimation of all the unknown parameters (i.e. $\alpha, \beta_1, \beta_2, \beta_3, \beta_4$), of key interest is the estimated β_4 , as this provides the adjusted prognostic effect of the index prognostic factor, revealing its independent contribution to the prediction of the outcome irrespective of the prognostic effects of the other (established) factors A_1 , A_2 and A_3 combined.

The need to focus on adjusted prognostic effects is no different from systematic reviews of aetiological studies, where the focus is on estimating the association of a certain causal risk factor after adjustment for other risk factors. In such causal research these factors are usually referred to as 'confounders' rather than as 'other prognostic factors' which is typically used for prognosis research. The crude (unadjusted) prognostic effect of some index factor may completely disappear after adjustment and is therefore rather uninformative, especially since prognostication in

healthcare is rarely based on a single prognostic factor but rather on the information of multiple prognostic factors.⁴

In this article our focus is on systematic reviews to summarise prognostic effect estimates. Some primary studies may also evaluate an index factor's added value in terms of improvement in risk classification and clinical utility (e.g. using measures such as net reclassification improvement and net-benefit), and change in prediction model performance (e.g. by calculating the change in the concordance index, also known as the C-statistic or area under the ROC curve).¹⁶⁻¹⁹ However, this is beyond the scope of this article, and we refer the reader to other relevant sources.^{6 20 21}

Application to the CRP review:

CRP is widely studied for its prognostic value in patients with coronary disease, but there is continued uncertainty as to whether it is useful as US and European clinical practice guidelines recommend measurement but clinical practice varies widely. This motivated the systematic review by Hemingway et al.,¹⁴ for which the corresponding PICOTS system is shown in Figure 1. No studies were excluded on the basis of methodological standards, sample size, duration of follow-up, publication year, or language of publication.

Step 2: Searching and selection of eligible studies

The next step is to identify primary studies that are eligible for the review; studies that address the review question defined in step 1 following the PICOTS framework. Unfortunately, it is more difficult to identify prognostic factor studies than randomised trials of interventions. Prognosis studies do not tend to be indexed ('tagged') because a taxonomy of prognosis research is not widely recognised. Moreover, compared to studies of interventions, in prognostic factor studies there is much more variation in designs (e.g. cohort study data, randomised trial data, routine care registry data and case control study data can all be used), patient inclusion criteria, predictor and outcome measurement, follow up time, methods of statistical analysis, and in the adjustment of (and number of) other prognostic factors (covariates). Between-study heterogeneity is thus the rule rather than the exception in prognostic factor research. It is therefore essential systematic reviews of prognostic factor studies define the study inclusion and exclusion criteria based on the PICOTS structure (step 1), as it determines the study search and selection strategy.

Typically broad search and selection filters are required, combining terms related to prognosis research (such as prognostic, predict, predictor, factor, independent) with domain or disease specific terms (such as the name of prognostic factors and the targeted disease or patient population).²² Such a broad search comes at the (often considerable) expense of retrieving many irrelevant records. Geersing et al.²³ validated various existing search strategies for prognosis studies and also suggested a generic filter for identifying studies of prognostic factors,^{22 24 25} which extended the work of Ingui, Haynes and Wong.^{22 24 25} When tested in a single review of prognostic factors, this generic filter had a number needed to read (NNR) of 569, emphasising the difficulty in targeting prognostic factor articles. The NNR might be considerably reduced in situations where specific factors and/or populations are added to the filter. Even then, care is still needed to be inclusive, as multiple terms are often used for the same meaning; for example, biomarker MYCN is also referred to as n-myc and nmyc amongst others.¹³

Once the search is complete, each potentially relevant study must be screened for their applicability to the review question. Due to the aforementioned heterogeneity in prognostic factor studies, during this study selection phase more deviations from the defined PICOTS (in step 1) are possible (indeed, far greater than what is typically encountered during the selection of randomised intervention studies). The applicability of this primary study selection should first be based on title and abstracts screening, followed by full text screening (see below) both ideally done by two researchers independently. Any discrepancies should be resolved through discussion, and potentially with a third reviewer. To check if any relevant articles have been missed, it is helpful to share the list of identified studies with researchers in the field, to examine the reference lists of identified articles, and to perform a citation search.

Figure 1 : Six items, abbreviated as PICOTS, to help define the question for systematic reviews of prognostic factor studies as based on the CHARMS guidance¹⁵ and applied to a review of the adjusted prognostic value of CRP.¹⁴

- **Population:** *define the target population in which the prognostic factor(s) under review are to be used.*
e.g. CRP review: patients with stable coronary disease, defined as clinically diagnosed angina pectoris or angiographic disease, or a history of previous acute coronary syndrome at least 2 weeks prior to prognostic factor (CRP) measurement.
- **Index prognostic factor:** *define the factor(s) whose prognostic value is under review.*
e.g. CRP review: CRP was the single biomarker reviewed for its prognostic value.
- **Comparator prognostic factor(s):** *comparator prognostic factors can be considered in a review in various ways. For example, the aim might be to compare the prognostic ability of a certain index factor to two or more other (i.e. comparator) prognostic factors. Also, the aim may be to review the adjusted prognostic value of a particular index factors, i.e. over and above (adjusted for, independent of) other existing (i.e. comparator) prognostic factors. If the only aim is to summarise the unadjusted prognostic effect of a particular index factor, then there is actually no comparator factor being addressed.*
e.g. CRP review: the focus was on the adjusted prognostic value of CRP; i.e. its prognostic effect after adjusting for existing (comparator) prognostic factors. In particular, adjustment for the following conventional prognostic factors was of interest: age, sex, smoking status, obesity, diabetes, and one or more lipid variables [from total cholesterol, LDL cholesterol, HDL cholesterol, triglycerides], and inflammatory markers [fibrinogen, IL-6, white cell count]).
- **Outcome:** *define the outcome(s) for which the prognostic ability of the factor(s) under review are of interest.*
e.g. CRP review: outcome events were defined as coronary (coronary death, sudden cardiac death, acute nonfatal myocardial infarction, primary percutaneous coronary intervention, unplanned emergency admissions with unstable angina), cardiovascular (where coronary events were reported in combination with heart failure, stroke, or peripheral arterial disease), and all-cause mortality.
- **Timing:** *define (i) at what time-point(s) the prognostic factors (index and comparators) are to be used (i.e. the time point of prognostication) and (ii) over what time period the outcome(s) are predicted by these factors.*
e.g. CRP review: there was no restriction on the time-points and time period. The CRP measurement had to be done at least two weeks after diagnosis and all follow-up information on the outcomes (all time periods) was extracted from the studies.
- **Setting:** *define the intended role or setting of the prognostic factor(s) under review.*
e.g. CRP review: CRP measurement was studied in both primary and secondary care to provide prognostic information about patients diagnosed with coronary heart disease, and thus may be useful for healthcare professionals treating and managing such patients.

Application to the CRP review:

Hemingway et al. included any prospective observational study that reported risk of subsequent events among patients with stable coronary disease in relation to measured CRP values.¹⁴ Eligible studies had to include patients with stable coronary disease, defined as clinically diagnosed angina pectoris or angiographic disease, or a history of previous acute coronary syndrome at least 2 weeks prior to CRP measurement. They searched MEDLINE between 1966 and 25 November 2009 and EMBASE between 1980 and 17 December 2009, using a search string containing terms for coronary disease, prognostic studies, and CRP. The search identified 1,566 articles of which 83 studies fulfilled the inclusion criteria. Had specific terms for CRP not been included in the search string, then the total number of identified articles would have far exceeded 1,566.

Step 3: Data extraction

The next step is to extract key information from each selected study. Data extraction provides the necessary data from each study, which enables reviewers to examine their (eventual) applicability and risk of bias (see step 4). It also provides the information required for subsequent qualitative and quantitative (meta-analysis) synthesis of the evidence across studies. The CHARMS checklist provides explicit guidance (see Table 2 in Moons et al¹⁵) about which key items across 11 domains should be extracted from primary studies of prediction models, and for what reason (i.e. to provide general information of the primary study, to guide risk of bias assessment or to judge applicability of the primary study for the review question at hand). In **Figure 2**, based on considerable experience of conducting systematic reviews of prognostic factor studies, we modified the original CHARMS checklist for prediction model studies, to make it suitable for data extraction purposes in reviews of prognostic factors; we refer to it as CHARMS-PF. This basically means that three domains typically addressing prediction modelling were combined to one overall Analysis domain, while other domain names and key items were slightly reworded or extended. Reasons for extraction of each key item were similar to the CHARMS checklist for prediction models. As we developed the original CHARMS checklist, a wider consensus agreement of the CHARMS-PF content was not considered necessary.

Reviewers should extract fundamental information from the studies, such as the dates, setting, study design, definitions of start-points, outcomes, follow-up length, and prognostic factors; reviewers will often find large heterogeneity across studies in these aspects. The extracted information allows for summary tables of study characteristics. In addition, more specific information is needed for proper risk of bias and additional applicability assessment (see step 4), such as methods of measurement of the prognostic factors and outcomes, the handling of missing

data, attrition (lost to follow-up), and whether estimated associations of the prognostic factors under review were adjusted for other prognostic factors. This also enhances the potential for meta-analysis and the presentation and interpretation of subsequent summary results (see steps 5-8).

To enable meta-analysis of prognostic factor studies, the key elements to extract are estimates, and corresponding standard errors or confidence intervals, of the prognostic effect for each factor of interest; for example, the estimated risk ratio or odds ratio (for binary outcomes), the hazard ratio (for time-to-event outcomes), or mean difference (for continuous outcomes). As most prognostic factor studies consider time-to-event outcomes (including censored observations and different follow-up lengths for patients), hazard ratios are often the most suitable effect measure. A concern is that hazard ratios may not be constant over time, and therefore any evaluations of non-proportional hazards (i.e. non-constant hazard ratios for the prognostic factors of interest) should also be extracted. Such information may unfortunately be rarely reported.

Unfortunately, many prognostic factor studies do not adequately report estimated effect measures and/or their precision. For this reason, methods are available to restore the missing information upon data extraction. In particular, Parmar et al.²⁶ and Tierney et al.²⁷ describe how to obtain unadjusted hazard ratio estimates (and their variances) when they are not reported directly. For example, under assumptions, the number of outcomes (events) and an available *p*-value (e.g. from a log-rank test or Cox regression) can be used to indirectly estimate the unadjusted hazard ratio between two groups defined by a particular factor (e.g. "positive" versus "negative" levels). Perneger et al.²⁸ suggest how to derive unadjusted hazard ratios from survival proportions, and Perez suggest to use a simulation approach.²⁹ Even with such indirect estimation methods, not all results will be obtainable. For example, in a systematic review of 575 studies investigating prognostic factors in neuroblastoma,³⁰ the methods of Parmar et al. were used to obtain 204 hazard ratios estimates and their confidence intervals; but this represented only 35.5% of the potential evidence.

Figure 2: The CHARMS-PF checklist of key items to be extracted from primary studies of prognostic factors, based on additions and modifications of the original CHARMS checklist for primary studies of prediction models¹⁵, to serve general description, assessment of applicability or risk of bias.

Domain	Key items	General	Applicability	Risk of Bias
SOURCE OF DATA	Source of data (e.g. cohort, case-control, randomised trial participants, or registry data)		X	X
PARTICIPANTS	Participant eligibility and recruitment method (e.g. consecutive participants, location, number of centres, setting, inclusion and exclusion criteria)	X	X	X
	Participants description	X	X	
	Details of treatments received, if relevant		X	X
	Study dates	X	X	
OUTCOME(S) TO BE PREDICTED	Definition and method for measurement of outcome(s)		X	X
	Was the same outcome definition (and method for measurement) used in all participants?			X
	Type of outcome(s) (e.g. single or combined endpoints)	X	X	
	Was the outcome(s) assessed without knowledge of the candidate prognostic factors (i.e. blinded)?			X
	Were candidate prognostic factors part of the outcome (e.g. when using a panel or consensus outcome measurement)?			X
	Time of outcome(s) occurrence or summary of duration of follow-up	X	X	X
PROGNOSTIC FACTORS (including index and comparator prognostic factors)	Number and type of prognostic factors (e.g. obtained from demographics, patient history, physical examination, additional testing, disease characteristics)	X		
	Definition and method for measurement of prognostic factors		X	X
	Timing of prognostic factor measurement (e.g. at patient presentation, at diagnosis, at treatment initiation, end of surgery)		X	X
	Were prognostic factors assessed blinded for outcome, and for each other (if relevant)?			X
	Handling of prognostic factors in the modelling (e.g. continuous, linear, non-linear transformations or categorised)			X
SAMPLE SIZE	Was a sample size calculation conducted and, if so, how?	X		
	Number of participants and number of outcomes/events	X		
	Number of outcomes/events in relation to the number of candidate prognostic factors			X
MISSING DATA	Number of participants with any missing value (in the prognostic factors and outcomes)	X		X
	Number of participants with missing data for each prognostic factor of interest			X
	Details of attrition (loss to follow-up) and, for time-to-event outcomes, number of censored observations (ideally in each category for those categorical prognostic factors of interest)			X
	Handling of missing data (e.g. complete-case analysis, imputation, or other methods)			X
ANALYSIS	Modelling method (e.g. linear, logistic, Cox, parametric survival, competing risks)	X		X
	How modelling assumptions were checked. In particular, for time-to-event outcomes and the analysis of hazard ratios, the method for assessing non-proportional hazards (non-constant hazard ratios over time).			X
	Method for selection of prognostic factors for inclusion in multivariable modelling (e.g. all candidate prognostic factors considered, pre-selection of established prognostic factors, retain only those significant from univariable analysis)			X
	Method for selection/exclusion of prognostic factors (including those of interest and those used as adjustment factors) during multivariable modelling (e.g. backward or forward selection, or full model approach including all factors regardless) and criteria used for any selection/exclusion (e.g. p-value, Akaike Information Criterion)			X
	Method of handling each continuous prognostic factor (e.g. dichotomisation, categorisation, linear, non-linear), including values of any cut-points used and their justification. For non-linear trends, the method of identifying non-linear relationships (e.g. splines, fractional polynomials).			X

RESULTS	Unadjusted and adjusted prognostic effect estimates (e.g. risk ratios, odds ratios, hazard ratios, mean differences) for each prognostic factor of interest, and the corresponding 95% confidence interval (or variance or standard error). Details of any non-linear relationships and whether modelling assumptions hold. In particular, for time-to-event outcomes, any evidence of non-proportional hazards (non-constant hazard ratios) for each prognostic factor of interest.	X	X	
	For each extracted adjusted prognostic effect estimate of interest, the set of adjustment factors used.	X	X	
INTERPRETATION AND DISCUSSION	Interpretation of presented results	X	X	
	Comparison with other studies, discussion of generalizability, strengths and limitations.	X	X	

Although indirect estimation methods help retrieve *unadjusted* prognostic effect estimates, they often have limited value for obtaining *adjusted* effect estimates. Furthermore, even when multiple studies do provide the adjusted prognostic effect of a particular factor, then the set of adjustment factors will usually differ across studies. This complicates the interpretation of subsequent meta-analysis results. We recommend that reviewers, therefore, pre-define the core set of prognostic factors for the outcome of interest (e.g. age, gender, smoking, disease stage, etc.), that represents the desired ‘minimal’ set of adjustment factors. A consensus process, amongst health professionals and researchers in the field, may be required to agree this set. For example, a list of established prognostic factors could be identified that are routinely used within current prognostication of the clinical population of interest.

It may also be necessary to standardise the extracted estimates, to ensure they all relate to the same scale and direction in each study. In particular, the direction of effect will need standardising if one study compares the hazard rate in a factor’s ‘high’ versus ‘normal’ group, whereas another study compares the hazard rate in the factor’s ‘normal’ group versus ‘high’ group. When the outcome is defined differently across studies, approaches to convert effect measures on different outcome scales might also be useful.³¹ To deal with different cut-point levels for a particular continuous factor factor’s values,³² which allows them to convert prognostic effects of ‘high’ versus ‘normal’ to prognostic effects relating to a 1-unit increase in the factor. A similar approach was used by Hemingway et al.¹⁴ A concern, however, is that the actual distribution of a prognostic factor may be unknown (or even vary across studies). Finally, it is also possible to derive standardized effect estimates by standardizing the corresponding regression coefficients.³³

Application to the CRP review:

Hemingway et al extracted background information such as year of study start, number of included patients, mean age, baseline coronary morbidity (e.g. proportion with stable angina), average levels of biomarker at baseline, method of CRP measurement, follow-up duration, and number and type of

events. Basic information was often missing. For example, nearly a fifth of studies did not report the method of measurement, and only a quarter gave the number of patients included in the analyses and reasons for dropout. Prognostic effect estimates for CRP were extracted in terms of either the reported risk ratio, odds ratio or hazard ratio (labelled as 'relative risk'), and their 95% confidence intervals. These effect estimates were then converted to a standardised scale comparing the highest third with the lowest third of the (log-transformed) CRP distribution. Where available, separate prognostic effect estimates were extracted for different degrees of adjustment for other prognostic factors.

Step 4: Evaluating applicability and risk of bias of primary studies

Once eligible studies have been identified and data has been extracted, an important next step is to assess the *applicability* and *risk of bias* (quality) of each study for the review. As for steps 2 and 3, ideally this is done by two reviewers, independently, with any discrepancies resolved. As mentioned, applicability refers to the extent to which a selected study (in step 2) matches the review question in terms of the population, timing, prognostic factors and outcomes (endpoints) of interest. Just because a study is eligible for inclusion does not mean it is free from applicability concerns. A study may be applicable in some aspects (e.g. correct condition at start-point, with prognostic factors of interest evaluated) but not others (e.g. incorrect population or setting, inappropriate outcome definition, too different follow-up time, lack of adjustment for conventional prognostic factors). Applicability is typically done at the title and abstract screening as well as in this step based on full text screening, obviously determined by the PICOTS (step 1) and in/exclusion criteria of studies (step 2).

Risk of bias refers to the extent to which flaws in the study design or analysis methods may lead to bias in estimates of the prognostic factor effects. Unfortunately, based on growing empirical evidence from systematic reviews examining methodology quality, many primary studies will be at high risk of bias.^{30 34-36 8 37-40 41 42} For prognostic factor studies, Hayden et al. developed the QUIPS checklist for examining risk of bias across six domains:⁴³ Study Participation, Study Attrition, Prognostic Factor Measurement, Outcome Measurement, Adjustment for other prognostic factors, and Statistical Analysis and Reporting. **Figure 3** shows the signalling items within these domains, to help guide reviewers toward low, unclear or high risk of bias classifications. Additional guidance may be found from general tools examining the quality of observational studies,^{44 45} and the REMARK guideline for reporting of primary prognostic factor studies.^{46 47}

We recommend that users first operationalise criteria to assess the signalling items and domains for the specific review question. For example, with the study participation and attrition domains, this includes defining a priori the most important characteristics that may indicate a systematic bias in study recruitment (study participation domain) and loss to follow-up (study attrition domain). Defining these characteristics ahead of time will facilitate assessment and consensus related to the importance of potential differences that could influence the observed association between the index prognostic factors and outcomes of interest. Definitions of sufficiently valid and reliable measurement of the index prognostic factor(s) and outcome(s) should also be specified at the protocol stage. Similarly the core set of adjustment prognostic factors, that are deemed necessary for the primary studies to have used, should be pre-defined to facilitate judgement related to risk of bias of domain 5.

Overall assessment of the six risk of bias domains is undertaken by considering the risk of bias information from the signalling items for each domain, rated as low, moderate and high risk of bias. Occasionally, item information to assess the bias domains is not available in the study report. It is then recommend consulting other publications that may have used the same dataset (as often in prognostic studies based on large existing cohorts), and contacting study authors for additional information. An informed judgement about the potential risk of bias for each bias domain should be made independently by two reviewers, and discussed to reach consensus. Each of the six domains needs to be rated and reported each of these to inform readers, inform the field for future primary studies, and to facilitate future meta-epidemiological research. To make a judgment about the overall study risk of bias, we recommend to describe studies with a 'low risk of bias' as those studies where all, or the most important domains (as determined a priori), are rated as having low (or low/moderate) risk of bias.

Application to the CRP review:

Hemingway et al. infer the quality of included studies by the quality of their reporting on 17 items derived from the REMARK guidelines.⁴⁷ The median number of study quality items reported was 7 out of a possible 17, and standards did not change between 1997 and 2009. Only two studies referred to a study protocol, with none referring to a statistical analysis plan. Hemingway et al. note that this "makes it difficult to know what the specific research objectives were at the start of cohort recruitment, at the time of CRP measurement, or at the onset of the statistical analysis."¹⁴ Only two studies reported the time elapsed between first lifetime presentation with coronary disease and assessment of CRP, raising applicability concerns.

Figure 3: The QUIPS tool to assess Quality in Prognostic factor Studies, which can be used to classify risk of bias of prognostic factor studies. We modified some wording of Hayden et al.,⁴³ to be consistent with terminology used in our article.

Domains	Signalling items	Ratings
1. Study Participation	<ul style="list-style-type: none"> a. Adequate participation in the study by eligible persons b. Description of the target population or population of interest c. Description of the baseline study sample d. Adequate description of the sampling frame and recruitment e. Adequate description of the period and place of recruitment f. Adequate description of inclusion and exclusion criteria 	<p>High bias: The relationship between the PF and outcome is very likely to be different for participants and eligible nonparticipants</p> <p>Moderate: The relationship between the PF and outcome may be different for participants and eligible nonparticipants</p> <p>Low bias: The relationship between the PF and outcome is unlikely to be different for participants and eligible nonparticipants</p>
2. Study Attrition	<ul style="list-style-type: none"> a. Adequate response rate for study participants b. Description of attempts to collect information on participants who dropped out c. Reasons for loss to follow-up are provided d. Adequate description of participants lost to follow-up e. There are no important differences between participants who completed the study and those who did not 	<p>High bias: The relationship between the PF and outcome is very likely to be different for completing and non-completing participants</p> <p>Moderate: The relationship between the PF and outcome may be different for completing and non-completing participants</p> <p>Low bias: The relationship between the PF and outcome is unlikely to be different for completing and non-completing participants</p>
3. Prognostic Factor Measurement	<ul style="list-style-type: none"> a. A clear definition or description of the PF is provided b. Method of PF measurement is adequately valid and reliable c. Continuous variables are reported or appropriate cut points are used d. The method and setting of measurement of PF is the same for all study participants e. Adequate proportion of the study sample has complete data for the PF f. Appropriate methods of imputation are used for missing PF data 	<p>High bias: The measurement of the PF is very likely to be different for different levels of the outcome of interest</p> <p>Moderate: The measurement of the PF may be different for different levels of the outcome of interest</p> <p>Low bias: The measurement of the PF is unlikely to be different for different levels of the outcome of interest</p>

4. Outcome Measurement	<p>a. A clear definition of the outcome is provided</p> <p>b. Method of outcome measurement used is adequately valid and reliable</p> <p>c. The method and setting of outcome measurement is the same for all study participants</p>	<p>High bias: The measurement of the outcome is very likely to be different related to the baseline level of the PF</p> <p>Moderate: The measurement of the outcome may be different related to the baseline level of the PF</p> <p>Low bias: The measurement of the outcome is unlikely to be different related to the baseline level of the PF</p>
5. Adjustment for other prognostic factors	<p>a. All other important PFs are measured</p> <p>b. Clear definitions of the important PFs measured are provided</p> <p>c. Measurement of all important PFs is adequately valid and reliable</p> <p>d. The method and setting of PF measurement are the same for all study participants</p> <p>e. Appropriate methods are used to deal with missing values of PFs, such as multiple imputation</p> <p>f. Important PFs are accounted for in the study design</p> <p>g. Important PFs are accounted for in the analysis</p>	<p>High bias: The observed effect of the PF on the outcome is very likely to be distorted by another factor related to PF and outcome</p> <p>Moderate: The observed effect of the PF on outcome may be distorted by another factor related to PF and outcome</p> <p>Low bias: The observed effect of the PF on outcome is unlikely to be distorted by another factor related to PF and outcome</p>
6. Statistical Analysis and Reporting	<p>a. Sufficient presentation of data to assess the adequacy of the analytic strategy</p> <p>b. Strategy for model building is appropriate and is based on a conceptual framework or model</p> <p>c. The selected statistical model is adequate for the design of the study</p> <p>d. There is no selective reporting of results</p>	<p>High bias: The reported results are very likely to be spurious or biased related to analysis or reporting</p> <p>Moderate: The reported results may be spurious or biased related to analysis or reporting</p> <p>Low bias: The reported results are unlikely to be spurious or biased related to analysis or reporting</p>

PF = prognostic factor.

Step 5: Meta-analysis

Meta-analysis of prognostic factor studies aim to summarise the (adjusted) prognostic effect of each factor of interest. Aside from missing estimates (discussed earlier), challenges for the meta-analyst include: (i) having different types of prognostic effect measures (e.g. odds ratio and hazard ratios) which are not necessarily comparable;²⁸ (ii) estimates without standard errors, a problem as meta-

analysis methods typically weight each study by (a function of) their standard error; (iii) estimates relating to different time-points of the outcome occurrence/measurement; (iv) different methods of measurement for prognostic factors and outcomes; (v) different sets of adjustment factors; and (vi) different approaches to handling continuous prognostic factors (e.g. categorisation, linear, non-linear trends), including the choice of cut-point value when dichotomising continuous values into 'high' and 'low' groups. Many of these issues lead to substantial heterogeneity, such that – if meta-analysis is performed - summary results then have no direct interpretation.

Generally, meta-analysis results will be most interpretable, and thus useful, when a separate meta-analysis is undertaken for groups of 'similar' prognostic effect measures. In particular, we suggest to consider meta-analysis for:

- Hazard ratios, odds ratios and risk ratios separately
- Unadjusted and adjusted associations separately
- Prognostic factor effects at distinct cut-points (or groups of similar cut-points) separately
- Prognostic factor effects corresponding to a linear trend (association) separately
- Prognostic factor effects corresponding to non-linear trends separately
- Each method of measurement (for factors and outcomes) separately

Furthermore, ideally a meta-analysis of adjusted results should ensure that all included estimates are adjusted for the same set of other prognostic factors. This is unlikely, and so a compromise could be to ensure that all adjusted estimates in the same meta-analysis have adjusted for at least a (pre-defined) minimum set of adjustment factors (i.e. established prognostic factors).

Even when adhering to this guidance, unexplained heterogeneity is likely to remain due to other factors (e.g. length of follow-up, treatments received during follow-up). Therefore, if meta-analysis is performed, a random effects approach is essential to allow for unexplained heterogeneity across studies (**Figure 4**), as previously described in the BMJ.⁴⁸ This provides a summary estimate of the average prognostic effect of the factor, and the variability in effect across studies. Also potentially useful are meta-analysis methods to estimate the trend (e.g. linear effect) of a prognostic factor that has been grouped into three or more categories within studies (with each category compared to the reference category). These methods generally model the estimated prognostic effect sizes in each category as a function of 'exposure' level (e.g. mid-point or median prognostic factor value in the category), whilst accounting for within-study correlation and between-study heterogeneity.⁴⁹⁻⁵³ To apply these methods, some additional knowledge of the factor's underlying distribution is usually needed to help define the 'exposure' level, as the chosen value can impact upon the results.⁵¹

Advanced multivariate meta-analysis methods are also available to jointly handle multiple cut-points,⁵⁴ multiple methods of measurement,⁵⁴ or different adjustment factors in prognostic factor studies.⁵⁵ An introduction to multivariate meta-analysis is provided previously in the BMJ.⁵⁶

Figure 4: Explanation of a random effects meta-analysis of prognostic factor effect estimates

The true prognostic effect of a factor is likely to vary from study to study, and thus assuming a common (fixed) prognostic effect is not sensible. If Y_i and $\text{var}(Y_i)$ denote the prognostic effect estimate (e.g. $\ln(\text{hazard ratio})$, $\ln(\text{odds ratio})$, $\ln(\text{risk ratio})$, or mean difference) and its variance in study i , then a general random effects meta-analysis model can be specified as:

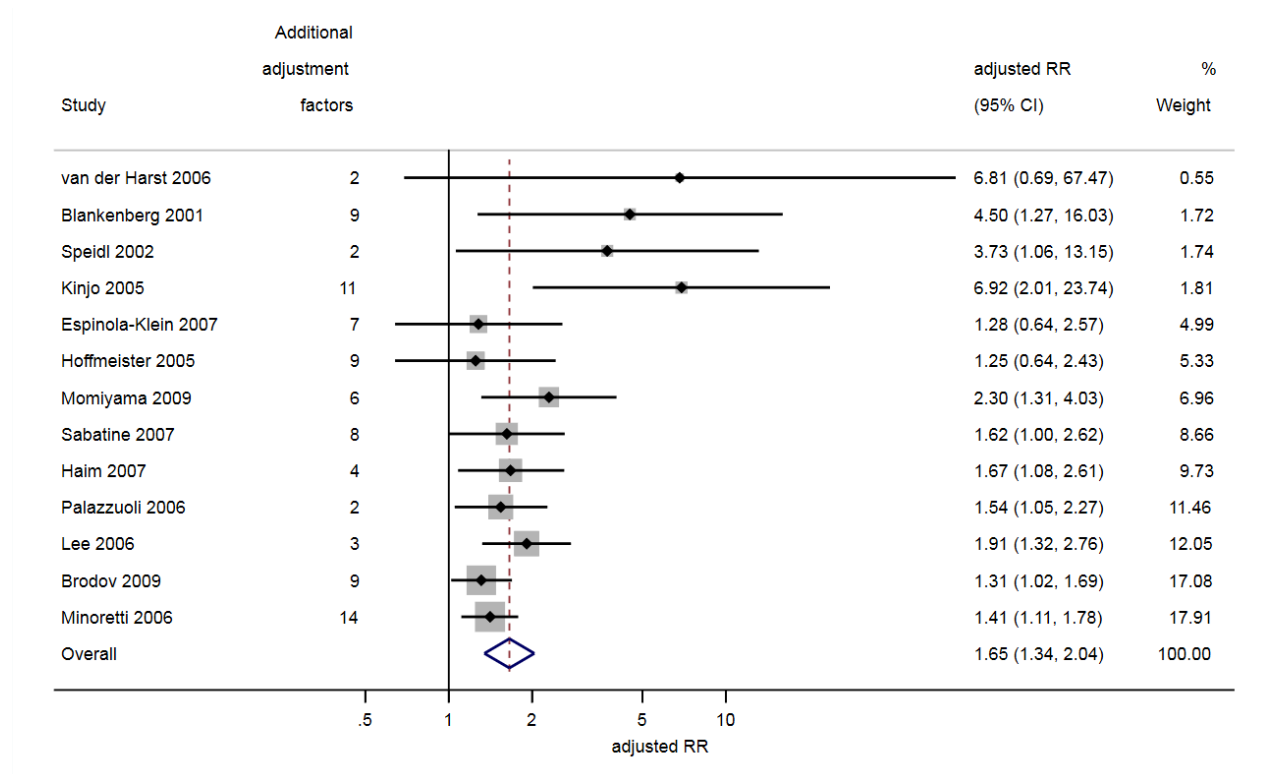
$$Y_i \sim N(\mu, \text{var}(Y_i) + \tau^2),$$

Most researchers use either restricted maximum likelihood or the approach of DerSimonian and Laird to estimate this model,⁵⁷ but other options are available including a Bayesian approach.⁵⁸ Of key interest is the summary (average) estimate, $\hat{\mu}$, which reveals the average prognostic effect of the factor. The standard deviation of the prognostic effect across studies is denoted by τ , and non-zero values suggest there is between-study heterogeneity. Confidence intervals for μ should ideally account for uncertainty in estimated variances (in particular τ),⁵⁹ and we have found the approach of Hartung-Knapp to be robust for this purpose in most settings.^{60 61} When synthesising prognostic effects on the log scale, the summary results and confidence intervals require back-transformation (using the exponential function) to the original scale.

Application to the CRP review:

Hemingway et al.¹⁴ apply a random effects meta-analysis to combine 53 adjusted prognostic effect estimates for CRP from studies that adjusted for at least one of six conventional risk factors (age, gender, smoking, diabetes, obesity, and lipids). The summary meta-analysis result was a relative risk of 1.97 (95% CI: 1.78 to 2.17), which gives the average prognostic effect of CRP (for those in the top versus bottom third of CRP distribution), and suggests larger CRP values are associated with higher risk. Although there was substantial between-study heterogeneity, nearly all estimates were in the same direction (i.e. relative risk > 1). When restricting meta-analysis to just the 13 studies that adjusted for at least all six conventional prognostic factors, the summary relative risk decreased to 1.65 (95% CI: 1.39 to 1.96) and the between-study heterogeneity reduced. Using the study-specific estimates shown Hemingway et al. we updated this meta-analysis (**Figure 5**), obtaining the same summary result but a wider confidence interval (1.34 to 2.04) via the Hartung-Knapp approach.⁶⁰

Figure 5: Forest plot showing the study-specific estimates and meta-analysis summary result of the adjusted prognostic effect (relative risk, RR) of CRP taken from the review of Hemingway et al¹⁴; all studies were adjusted for age, gender, smoking, diabetes, obesity, and lipids, plus up to 14 other variables. Meta-analysis results shown are based on a random effects meta-analysis model with DerSimonian and Laird estimation of the between-study variances. The summary result is identical to Hemingway et al.,¹⁴ but the confidence interval is wider as, here, we used the Hartung-Knapp approach to account for uncertainty in variance estimates.⁶⁰



Step 6: Quantifying and examining heterogeneity

As applies to all meta-analyses, in situations with large heterogeneity across included studies, it may be better to refrain from synthesising the study results, and rather display the variability in estimates on a forest plot without showing an overall pooled estimate. When meta-analysis is still performed in the face of heterogeneity, it is important to quantify and report the magnitude of heterogeneity itself, for example via the estimate of τ^2 (the between-study variance),⁶² or an approximate 95% prediction interval indicating the potential true prognostic effect of a factor in a new population.^{48 63}

Subgroup analyses and meta-regression can be used to examine or explore the causes of heterogeneity. A subgroup analysis performs a separate meta-analysis for categories defined by a particular characteristic, such as those with low risk of bias, those with a follow-up < 1 year or ≥ 1 year, or those set in countries within Europe. A better approach is meta-regression, which extends the meta-analysis equation shown in **Figure 4** by including study-level covariates,⁶⁴ and allows a formal comparison of how meta-analysis results differ across groups defined by covariates (e.g. low

risk of bias studies versus other studies at higher risk of bias). Unfortunately, subgroup analyses and meta-regression are often problematic. There will often be few studies per subgroup and low power to detect genuine causes of heterogeneity. Furthermore, study-level confounding will be rife, such that it is difficult to disentangle the associations for one covariate from another. For example, studies with a low risk of bias may also have a different length of follow-up, or a particular cut-point level, compared with studies at higher risk of bias.

Application to the CRP review:

Hemingway et al. report that meta-regression identified four study-level covariates that explained some between-study heterogeneity in the prognostic effect of CRP: definition of comparison group, number of adjustment variables, the (log) number of events and the proportion of patients with stable coronary disease (both reflecting study size).¹⁴ Studies originally reporting unequal CRP groups had stronger effects than those reporting CRP on a continuous scale. For each additional adjustment factor the summary relative risk decreased by 3%. The summary relative risk was smaller among studies with more than the median number of outcome events, and smaller among studies confined to stable coronary disease. There was no evidence that the CRP effect differed according to the number of quality items reported by a study.

Step 7: Examining small-study effects

The term 'small-study effects' refers to when there is a systematic difference in prognostic effect estimates for small studies and large studies.⁶⁵ A particular concern is when small studies (especially those that are exploratory, as these often evaluate many potential prognostic factors with relative few outcome events) show larger prognostic effects than larger studies. This may be due to chance or heterogeneity, but a major threat is publication bias and selective reporting, which are endemic in prognosis research.^{26 62} Such reporting biases lead to those smaller studies with (statistically) significant or larger prognostic factor effect estimates more likely to be published or reported in sufficient detail, and thus included in meta-analysis, than those smaller studies with non-significant or smaller prognostic effect estimates. This is a potential concern for both unadjusted and adjusted prognostic effects. A primary study usually estimates an unadjusted prognostic effect for each of multiple prognostic factors, but then study authors may only report those that are statistically significant. Adjusted results are also often only reported for those prognostic factors that retain statistical significance in both univariable and multivariable analysis. A consequence is that meta-

analysis results will be biased, with larger summary prognostic effects than in truth, and potentially some factors being deemed to have clinical value when they actually do not.

The evidence for small-study effects is usually considered on a funnel plot, which shows the study estimates (x-axis) against their precision (y-axis). This is usually recommended if there are 10 or more studies.⁶⁵ This should ideally show a symmetric, funnel like shape, with results from larger studies at the centre of the funnel, with smaller studies spanning out in both directions equally. Asymmetry will arise if there are small-study effects, with a greater proportion of smaller studies in one particular direction. Statistical tests for asymmetry in risk, odds and hazard ratios can be used, such as Peter's and Debray's test.^{66 67} Contour-enhanced funnel plots also show the statistical significance of individual studies, and 'missing' studies are perhaps more likely to fall within regions of non-significance if publication bias was the cause of small-study effects.

However, as small-study effects may also arise due to heterogeneity, it is difficult to disentangle publication bias from heterogeneity in a single review. For example, if smaller studies used an analysis with fewer adjustment factors, then this may cause larger prognostic factor effects in such studies rather than due to publication bias. Multivariate meta-analysis might be able to reduce the impact of small study effects, by 'borrowing strength' from related information.⁵⁶

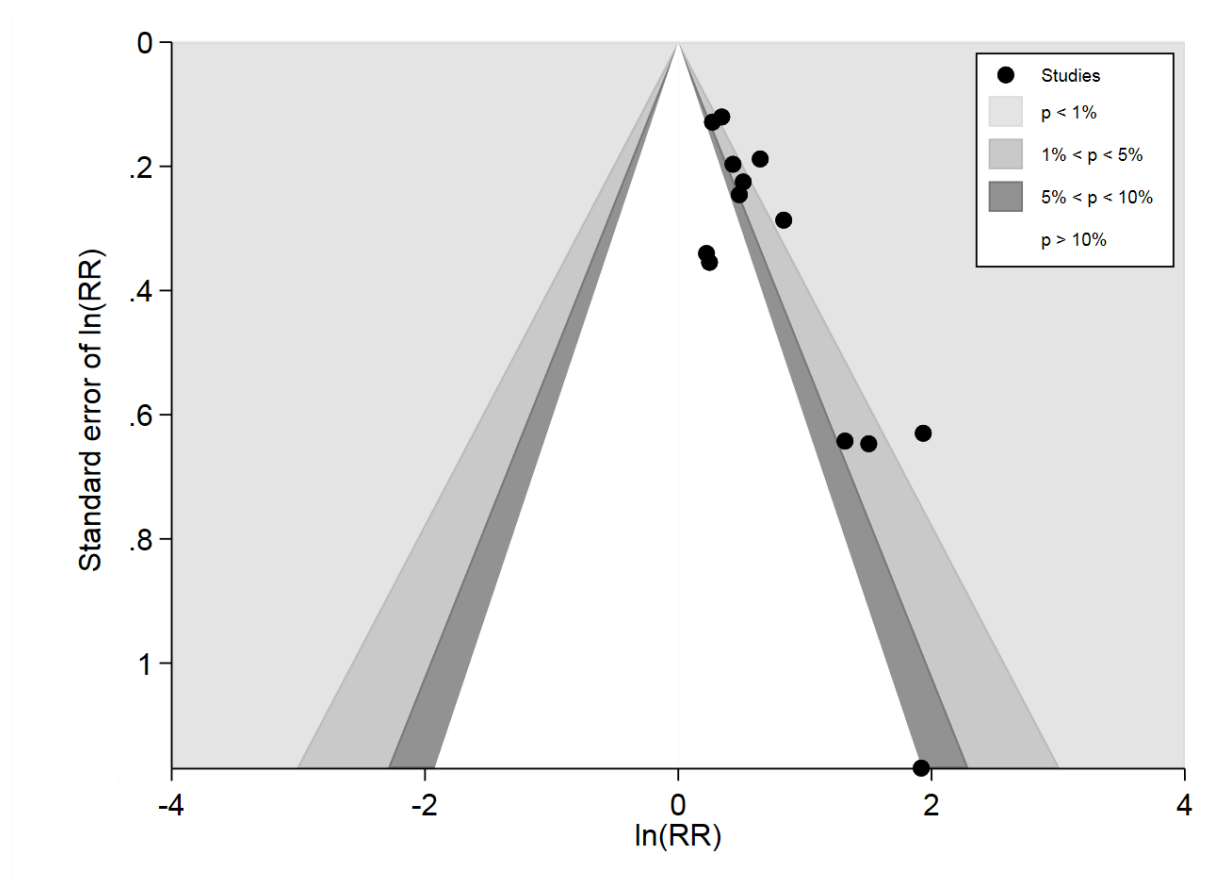
As previously mentioned, a related concern is that smaller prognostic factor studies are generally at higher risk of bias than larger studies. Smaller studies tend to be more exploratory in nature and typically based on a convenient sample, often examining many (sometimes hundreds) of potential prognostic factors, with relatively few outcome events. This leads to spurious (due to chance) and potentially biased (due to poor estimation properties⁶⁸) prognostic effect estimates. In contrast, larger studies are often confirmatory studies focusing on one or a few prognostic factors, and are more likely to adopt a protocol-driven and prospective approach, with clearer reporting regardless of their findings.³ As such, larger studies are less likely to identify spurious prognostic factor effect estimates. Therefore, it is helpful to examine small study effects (potential publication bias) when restricting analysis to the subset of studies at low of risk of bias. If this resolves previous issues of small study effects in the full meta-analysis, then it gives even more credence to focus conclusions and recommendations on the meta-analysis results based on just the higher quality studies.

Application to the CRP review:

Figure 6 shows a funnel plot of the study estimates within the CRP meta-analysis shown in **Figure 5**. There is clear asymmetry, raising a strong concern of publication bias. There was an insufficient

number of studies considered at low risk of bias to evaluate small-study effects in a subset of higher quality studies.

Figure 6: Evidence of funnel plot asymmetry (small-study effects) in the CRP meta-analysis shown in Figure 5. The smaller studies (those with higher standard errors) have relative risk (RR) estimates mainly to the right of the larger studies, and thus give the largest prognostic effect estimates. A concern is that this is due to publication bias, with 'missing' studies potentially falling in the white area denoting non-significant RR estimates.



Step 8: Reporting and interpretation of results

As with all research studies, clear and complete reporting is essential for reviews of prognostic factor studies. Most of the reporting guidelines of PRISMA and MOOSE will be relevant,^{69 70} and should be complemented by REMARK,^{46 47} which was aimed at primary prognostic factor studies. More specific guidance for reporting systematic reviews of prognostic factor studies is under development.

Interpretation and translation of summary meta-analysis results is an important final step. The guidance in the previous steps are the essential input for this. Discussion is necessary regarding if

and how the prognostic factors identified may be useful in practice (i.e. translation of results to clinical practice), and what further research is necessary. Ideally impact studies (e.g. randomised trials which compare groups which do and do not utilise a prognostic factor to inform clinical practice) are needed before strong recommendations for clinical practice are made; however, these are rare and outside the scope of the review framework outlined in this article.

Further, for interpreting the *certainty* (confidence) of the results of a review of intervention effectiveness, Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) was developed. This approach assesses the overall quality and certainty of evidence for the summary estimates of the intervention effects by addressing five domains: risk of bias, inconsistency, imprecision, indirectness and publication bias. The GRADE domains are judged using the information obtained by the tools and methods addressed in the above steps. Whether these domains, developed for reviews of interventions, are equally applicable to judge the certainty of summary results of systematic reviews of prognostic factor studies is yet unknown. Compared to reviews of intervention studies, allowing for heterogeneity (the inconsistency domain) might be more acceptable in reviews of prognostic factor studies due to the inevitable heterogeneity caused by study differences in methods of measurement, adjustment factors and statistical analysis methods, amongst others. Further, the threat of selective reporting or publication bias in reviews of prognostic factor studies may be more severe than in reviews of intervention studies, due to the aforementioned problems of exploratory studies, poor reporting, and biased analysis methods.

There is yet limited empirical evidence for the use of the existing domains for grading the certainty of summary estimates of prognostic factor studies, although a first attempt has been made,⁷¹ and one for grading the certainty of evidence of summary estimates of overall prognosis studies.⁷² We note that reviewers need to be especially cautious when comparing the adjusted prognostic value of different index factors, for example, to conclude whether the summary adjusted hazard ratio of prognostic factor A is larger than that for factor B. Usually different sets of studies will be available for each index factor, and so the comparison will be indirect and potentially biased. Moreover, the studies evaluating factor A may often have used different sets of adjustment factors (other prognostic factors) than those evaluating factor B. It will be rare to find studies on different index factors that used exactly the same set of adjustment factors. We therefore recommend reviewers restrict comparisons (of the adjusted prognostic value) of two or more index factors to those studies that at least used a similar, minimally required set of adjustment factors.⁷³

Application to CRP review:

The meta-analysis results suggest CRP is a prognostic factor for the risk of death and nonfatal cardiovascular events, even when only including the largest studies that adjusted for all six conventional prognostic factors. In their discussion, Hemingway et al. downgrade the meta-analysis findings, due to a strong concern about the quality and reliability of the underlying evidence.¹⁴ The absence of pre-specified protocols, poor and potentially biased reported, and strong potential for publication bias prevented them from making firm conclusions about whether CRP has prognostic factor after adjustment for established prognostic factors. They state that the concerns “explicitly challenge the statement for healthcare professionals made by the Centers for Disease Control that measuring CRP is both ‘useful’ and ‘independent’ as a marker of prognosis”.⁷⁴

Summary

We described the key steps and methods for conducting a systematic review and meta-analysis of prognostic factor studies. Current reviews are often limited by the quality and heterogeneity of primary studies.⁷⁵ We expect the prevalence of such reviews to grow rapidly, especially with Cochrane (via the Cochrane Prognosis Methods Group: www.methods.cochrane.org/prognosis/) currently embarking upon them. Our guidance will therefore help reviewers to write grant applications for reviews of prognostic factor studies, and to develop protocols of the review.⁷⁶ Such protocols should be published, ideally at the same time as the review is registered, for example within PROSPERO, the international prospective register of systematic reviews (www.crd.york.ac.uk/PROSPERO/), or the Cochrane database. Lastly, we note that some of the limitations described (e.g. use of different cut-point values across studies) could be alleviated if the individual participant data (IPD) were obtained from primary studies,⁷⁷ rather than being reliant on results extracted from study publications. Nevertheless, it cannot resolve all problems (e.g. quality of original study, availability of different adjustment factors), and it may take a number of years to obtain IPD from relevant studies.⁷⁸ Further discussion on IPD meta-analysis of prognostic factor studies is given elsewhere.⁷⁹

Competing interests: None

Exclusive licence: *The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above.*

Funding: Richard Riley and Kym Snell are supported by funding from the Evidence Synthesis Working Group, which is funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR) [ProjectNumber 390]. Kym Snell is also supported by a launching fellowship from the NIHR SPCR. Thomas Debray was supported by funding from the Netherlands Organisation for Health Research and Development (91617050 and 91215058). The views expressed are those of the author(s) and not necessarily those of the NIHR, the NHS, the Department of Health or the Netherlands Organisation for Health Research and Development. GSC was supported by the NIHR Biomedical Research Centre, Oxford

Contribution statement: RR and KM conceived the article content and structure. RR and KM wrote the first draft, and RR, KS, JE and TS added application to the CRP review. All authors provided intellectual content, text, and corrections to improve the first draft. KM, DA and GC co-developed the CHARMS guidance that informed the content of this paper. JH co-developed the QUIPS checklist, and informed the use and interpretation of QUIPS for this paper. RR, KM, and JH and subsequently all other authors revised the article after comments received by reviewers and the BMJ. RR is the guarantor.

Acknowledgements: We thank the BMJ Editors and three reviewers for their helpful feedback that improved the article upon revision.

Reference List

1. Leeflang MMG, Deeks JJ, Gatsonis C, et al. Systematic Reviews of Diagnostic Test Accuracy *Ann Intern Med* 2008;149:889-97.
2. Hemingway H, Croft P, Perel P, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ* 2013;346:e5595.
3. Riley RD, Hayden JA, Steyerberg EW, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med* 2013;10(2):e1001380.
4. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10(2):e1001381.
5. Hingorani AD, Windt DA, Riley RD, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ* 2013;346:e5793.
6. Debray TP, Damen JA, Snell KI, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;356:i6460.
7. Debray TPA, Damen JAAG, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res* 2018 (in-press).
8. Bouwmeester W, Zuithoff NP, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9(5):1-12.
9. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.
10. Altman DG. Systematic reviews of evaluations of prognostic variables. *BMJ* 2001;323(7306):224-8.
11. Altman DG, Riley RD. An evidence-based approach to prognostic markers. *Nature Clinical Practice Oncology* 2005;2:466-72.
12. Riley RD, Sauerbrei W, Altman DG. Prognostic markers in cancer: the evolution of evidence from single studies to meta-analysis, and beyond. *Br J Cancer* 2009;100(8):1219-29.
13. Riley RD, Burchill SA, Abrams KR, et al. A systematic review and evaluation of the use of tumour markers in paediatric oncology: Ewing's sarcoma and neuroblastoma. *Health Technol Assess* 2003;7(5):1-162.
14. Hemingway H, Philipson P, Chen R, et al. Evaluating the quality of research into a single prognostic biomarker: a systematic review and meta-analysis of 83 studies of C-Reactive protein in stable coronary artery disease. *PLoS Med* 2010;7(6):e1000286.
15. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11(10):e1001744.
16. Royston P. Explained variation for survival models. *Stata Journal* 2006;6:83-96.
17. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;23(5):723-48.
18. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016;352:i6.

19. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27(2):157-72.
20. Pennells L, Kaptoge S, White IR, et al. Assessing risk prediction models using individual participant data from multiple studies. *Am J Epidemiol* 2014;179(5):621-32.
21. Wynants L, Riley RD, Timmerman D, et al. Random-effects meta-analysis of the clinical utility of tests and prediction models. *Stat Med* 2018;37(12):2034-52.
22. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc* 2001;8(4):391-7.
23. Geersing GJ, Bouwmeester W, Zuithoff P, et al. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. *PLoS One* 2012;7(2):e32844.
24. Haynes RB, McKibbon KA, Wilczynski NL, et al. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ* 2005;330(7501):1179.
25. Wong SS, Wilczynski NL, Haynes RB, et al. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. *AMIA Annu Symp Proc* 2003:728-32.
26. Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat Med* 1998;17(24):2815-34.
27. Tierney JF, Stewart LA, Gherzi D, et al. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007;8:16.
28. Perner TV. Estimating the relative hazard by the ratio of logarithms of event-free proportions. *Contemporary Clinical Trials* 2008;29:762–66.
29. Perez T, McLellan J, Perera R. Extraction of unadjusted estimates of prognostic association for meta-analysis: simulation methods as good alternatives to trend and direct method estimation. *J Clin Epidemiol* 2018;99:153-63.
30. Riley RD, 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(8):1191-8.
31. Borenstein M, Hedges LV, Higgins JPT, et al. Converting Among Effect Sizes. In: Borenstein M, Hedges LV, Higgins JPT, et al., eds. *Introduction to Meta-Analysis*. West Sussex, UK.: John Wiley & Sons 2009.
32. Sadashima E, Hattori S, Takahashi K. Meta-analysis of prognostic studies for a biomarker with a study-specific cutoff value. *Research Synthesis Methods* 2016;7(4):402-19.
33. Nieminen P, Lehtiniemi H, Vähäkangas K, et al. Standardised regression coefficient as an effect size index in summarising the reported findings between quantitative exposure and response variables in epidemiological studies. *Epidemiology, Biostatistics and Public Health* 2013 10:e 8 8 5 4.
34. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. *European journal of cancer* 2007;43(17):2559-79.
35. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Quality of reporting of cancer prognostic marker studies: association with reported prognostic effect. *J Natl Cancer Inst* 2007;99(3):236-43.
36. Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. *J Natl Cancer Inst* 2005;97(14):1043-55.

37. Mallett S, Royston P, Dutton S, et al. Reporting methods in studies developing prognostic models in cancer: a review. *BMC medicine* 2010;8:20.
38. Collins GS, Mallett S, Omar O, et al. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine* 2011;9:103.
39. Collins GS, Omar O, Shanyinde M, et al. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013;66(3):268-77.
40. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *Br J Cancer* 2004;91(1):4-8.
41. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;14:40.
42. Mallett S, Royston P, Waters R, et al. Reporting performance of prognostic models in cancer: a review. *BMC medicine* 2010;8:21.
43. Hayden JA, van der Windt DA, Cartwright JL, et al. Assessing bias in studies of prognostic factors. *Ann Intern Med* 2013;158(4):280-6.
44. Sterne JA, Hernan MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919.
45. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomized studies in meta-analyses. 2009
[http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm].
46. Altman DG, McShane LM, Sauerbrei W, et al. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. *PLoS Med* 2012;9(5):e1001216.
47. McShane LM, Altman DG, Sauerbrei W, et al. REporting recommendations for tumour MARKer prognostic studies (REMARK). *Br J Cancer* 2005;93(4):387-91.
48. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011;342:d549.
49. Berlin JA, Longnecker MP, Greenland S. Meta-analysis of epidemiologic dose-response data. *Epidemiology* 1993;4(3):218-28.
50. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *Am J Epidemiol* 1992;135(11):1301-9.
51. Hartemink N, Boshuizen HC, Nagelkerke NJ, et al. Combining risk estimates from observational studies with different exposure cutpoints: a meta-analysis on body mass index and diabetes type 2. *Am J Epidemiol* 2006;163(11):1042-52.
52. Shi JQ, Copas JB. Meta-analysis for trend estimation. *Stat Med* 2004;23(1):3-19; discussion 159-62.
53. Orsini N, Li R, Wolk A, et al. Meta-analysis for linear and nonlinear dose-response relations: examples, an evaluation of approximations, and software. *Am J Epidemiol* 2012;175(1):66-73.
54. Riley RD, Elia EG, Malin G, et al. Multivariate meta-analysis of prognostic factor studies with multiple cut-points and/or methods of measurement. *Stat Med* 2015;34(17):2481-96.
55. Collaboration FS. Systematically missing confounders in individual participant data meta-analysis of observational cohort studies. *Stat Med* 2009;28(8):1218-37.

56. Riley RD, Jackson D, Salanti G, et al. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *BMJ* 2017;358:j3932.
57. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7.
58. Langan D, Higgins JP, Simmonds M. An empirical comparison of heterogeneity variance estimators in 12 894 meta-analyses. *Res Synth Methods* 2015;6(2):195-205.
59. Cornell JE, Mulrow CD, Localio R, et al. Random-effects meta-analysis of inconsistent effects: a time for change. *Ann Intern Med* 2014;160(4):267-70.
60. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med* 2001;20(24):3875-89.
61. Partlett C, Riley RD. Random effects meta-analysis: Coverage performance of 95% confidence and prediction intervals following REML estimation. *Stat Med* 2016.
62. Rucker G, Schwarzer G, Carpenter JR, et al. Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Med Res Methodol* 2008;8:79.
63. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society, Series A* 2009;172:137-59.
64. Berkey CS, Hoaglin DC, Mosteller F, et al. A random-effects regression model for meta-analysis. *Stat Med* 1995;14(4):395-411.
65. Sterne JAC, 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.
66. Debray TPA, Moons KGM, Riley RD. Detecting small-study effects and funnel plot asymmetry in meta-analysis of survival data: A comparison of new and existing tests. *Res Synth Methods* 2018;9(1):41-50.
67. Peters JL, Sutton AJ, Jones DR, et al. Comparison of two methods to detect publication bias in meta-analysis. *Jama* 2006;295(6):676-80.
68. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993;80(1):27-38.
69. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.
70. Stroup DF, Berlin JA, Morton SC, 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. Huguet A, Hayden JA, Stinson J, et al. Judging the quality of evidence in reviews of prognostic factor research: adapting the GRADE framework. *Systematic reviews* 2013;2:71.
72. Iorio A, Spencer FA, Falavigna M, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. *BMJ* 2015;350:h870.
73. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Ann Intern Med* 2013;158(7):544-54.
74. Pearson TA, Mensah GA, Alexander RW, et al. Markers of inflammation and cardiovascular disease: application to clinical and public health practice: A statement for healthcare professionals from the Centers for Disease Control and Prevention and the American Heart Association. *Circulation* 2003;107(3):499-511.

75. Sauerbrei W, Holländer N, Riley RD, et al. Evidence-based assessment and application of prognostic markers: the long way from single studies to meta-analysis. *Communications in Statistics* 2006;35:1333-42.
76. Peat G, Riley RD, Croft P, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. *PLoS Med* 2014;11(7):e1001671.
77. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010;340:c221.
78. Altman DG, Trivella M, Pezzella F, et al. Systematic review of multiple studies of prognosis: the feasibility of obtaining individual patient data. In: Auget J-L, Balakrishnan N, Mesbah M, et al., eds. *Advances in statistical methods for the health sciences* Boston: Birkhäuser 2006:3-18.
79. Abo-Zaid G, Sauerbrei W, Riley RD. Individual participant data meta-analysis of prognostic factor studies: state of the art? *BMC Med Res Methodol* 2012;12:56.