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

Richard D Riley,[1*] Karel G M Moons,[2,4*] Kym I E Snell,[1] Joie Ensor,[1] Lotty Hooft,[2,4] Douglas G Altman,[3] Jill Hayden,[5] Gary S Collins,[3] Thomas P A Debray[2,4]

[1]Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Keele, Staffordshire, ST5 5BG, UK
[2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
[3]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
[4]Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
[5]Centre for Clinical Research, Halifax, Nova Scotia, Canada
*Contributed equally

Correspondence to: R D Riley
r.riley@keele.ac.uk
(or @Richard_D_Riley on Twitter)

Cite this as: *BMJ* 2019;364:k4597
http://dx.doi.org/10.1136/bmj.k4597

Accepted: 8 October 2018

Prognostic factors are associated with the risk of future health outcomes in individuals with a particular health condition or some clinical start point (eg, a particular diagnosis). Research to identify genuine prognostic factors is important because these factors can help improve risk stratification, treatment, and lifestyle decisions, and the design of randomised trials. Although thousands of prognostic factor studies are published each year, often they are of variable quality and the findings are inconsistent. Systematic reviews and meta-analyses are therefore needed that summarise the evidence about the prognostic value of particular factors. In this article, the key steps involved in this review process are described.

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 the accuracy of screening and diagnostic tests,[1] the causal association of risk factors for disease onset, and the prognostic ability of bespoke factors and biomarkers. Prognostic evidence arises from prognosis studies, which aim to examine and predict future outcomes (such as death, disease progression, side effects or medical complications like pre-eclampsia) in people with a particular health condition or start point (such as those developing a certain disease, undergoing surgery, or women who are pregnant).

The PROGRESS (PROGnosis RESearch Strategy) framework defines four types of prognosis research objectives: (a) to summarise overall prognosis (eg, overall risk or rate) of health outcomes for groups with a particular health condition[2]; (b) to identify prognostic factors associated with changes in health outcomes[3]; (c) to develop, validate, and examine the impact of prognostic models for individualised prediction of such outcomes[4]; and (d) to identify predictors of an individual's response to treatment.[5] Each objective

SUMMARY POINTS

- Primary studies to identify prognostic factors are abundant, but often findings are inconsistent and quality is variable. Systematic reviews and meta-analyses are urgently needed to identify, evaluate, and summarise prognostic factor studies and their findings.
- A clear review question should be defined using the PICOTS system (Population, Index prognostic factor, Comparator prognostic factors, Outcome, Timing, Setting), 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 CHARMS (checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies) can be used for prognostic factors (CHARMS-PF).
- The QUIPS tool (quality in prognostic factor studies) can be used to examine each study's risk of bias. Unfortunately, many primary studies may have a high risk of bias because of poor design standards, conduct, and analysis. Applicability of a study should also be checked.
- If appropriate, meta-analysis can be used to combine prognostic effect estimates (such as hazard ratios, risk 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 prognostic effect estimates; adjusted estimates are important to examine a factor's independent prognostic value over and above (that is, after adjustment for) other 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 (such as hazard ratios or odds ratios).
- Publication bias is a major threat to the validity of meta-analyses of prognostic factor studies based on published evidence, and may cause small-study effects (asymmetry on a funnel plot).
- REMARK (reporting recommendations for tumour marker prognostic studies) and PRISMA (preferred reporting items for systematic reviews and meta-analyses) can be used to guide the reporting of the systematic review and meta-analysis of prognostic factor studies; the degree of confidence in the summary results from the review may be examined by use of adapted forms of GRADE (grades of recommendation, assessment, development, and evaluation) for interventions and diagnostic test accuracy studies.
- Availability of individual participant data from primary prognostic factor studies may alleviate many of the challenges.

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] In this article, we focus on prognostic factors.

A prognostic factor is any variable that is associated with the risk of a subsequent health outcome among people with a particular health condition. Different values or categories of a prognostic factor are associated with a better or worse prognosis of future health outcomes. For example, in many cancers, tumour grade at the time of histological examination is a prognostic factor because it is associated with time to disease recurrence or death. Each grade represents a group of patients with a different 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, comorbidities, 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 and lifestyle 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] Although thousands of such studies are published each year, often they are of variable quality and have inconsistent findings. Systematic reviews and meta-analyses are therefore urgently needed to summarise the evidence about the prognostic value of particular factors.[10-12] In this article, we provide a step-by-step guide on conducting such reviews. Our aim is to help readers, healthcare providers, and researchers understand the key principles, methods, and challenges of reviews of prognostic factor studies.

Step 1: Defining the review question

The first step is to define the review question. A review of prognostic factor studies falls within the second objective of the PROGRESS framework[2] because it aims to summarise the prognostic value of a particular factor (or each of multiple factors) for relevant health outcomes and time points in people with a specific health condition (eg, disease). Some reviews are broad; for example, Riley and colleagues aimed to identify any prognostic factor for overall and disease free survival in children with neuroblastoma or Ewing's sarcoma.[13] Other reviews have a narrower focus; for example, Hemingway and colleagues aimed to summarise the evidence on whether C reactive protein (CRP) is a prognostic factor for fatal and non-fatal events in patients with stable coronary disease.[14] This CRP review is used as an example throughout this article.

CHARMS (checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies) provides guidance for formulating a review question (table 1 in the article by Moons and colleagues[15]). Although CHARMS was developed[15] and 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. CHARMS[15] and subsequent improvements[6] propose a modification of the traditional PICO system (Population, Index intervention, Comparison and Outcome) used in systematic reviews of therapeutic intervention studies. The modification is called PICOTS, because it also considers Timing and Setting (box 1). In the context of prognostic factor reviews, the "P" of population and the "O" of outcome remain largely the same as in the original PICO system, but the "I" refers to index prognostic factors and the "C" refers to other prognostic factors that can be considered as comparators in some way. For example, the aim may be to compare the prognostic ability of a certain index factor with one or more other (that is, comparator) prognostic factors; or to investigate the adjusted prognostic value of a particular index factor over and above (adjusted for) other (that is, comparator) prognostic factors. If the only aim is to summarise the unadjusted prognostic effect of a particular index factor, which is not generally recommended, then no comparator factor is being considered. The "T" denotes timing and refers to two concepts of time. Firstly, at what time point the prognostic factors under review are to be measured or assessed (that is, the time point at which prognosis information is required); and secondly, over what time period the outcomes are predicted by these factors. The "S" of setting refers to the setting or context in which the index prognostic factors are to be used because the prognostic ability of a factor may change across healthcare settings.

An important component of reviews of prognostic factors is whether unadjusted or adjusted estimates of the index prognostic factors will be summarised, or both. We recommend that reviewers primarily focus on adjusted prognostic factor estimates because they reveal whether a certain index factor contributes independently to the prediction of the outcome over and above (that is, after adjustment for) other prognostic factors. In particular, for each clinical scenario there are often 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 the routinely measured ones. This means that reviewers need adjusted (and not unadjusted or crude) prognostic effect estimates to be estimated and reported in primary prognostic factor studies. Such adjusted prognostic estimates are typically derived from a multivariable regression model containing the established prognostic factors plus each index prognostic factor of interest.

Table 1 | 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 models[15]

Domain and key items	General	Applicability	Risk of bias
Source of data:			
Source of data (eg, cohort, case control, randomised trial, or registry data)	X	X	X
Participants:			
Participant eligibility and recruitment method (eg, consecutive participants, location, number of centres, setting, inclusion and exclusion criteria)	X	X	X
Participant description	X	X	
Details of treatments received (if relevant)		X	X
Study dates	X	X	
Outcomes to be predicted:			
Definition and method for measurement of outcomes		X	X
Was the same outcome definition (and method for measurement) used in all participants?			X
Types of outcomes (eg, single or combined endpoints)?	X	X	
Were the outcomes assessed without knowledge of the candidate prognostic factors (that is, blinded)?			X
Were candidate prognostic factors part of the outcome (eg, when using a panel or consensus outcome measurement)?			X
Time of outcome occurrence or summary of duration of follow-up	X	X	X
Prognostic factors (index and comparator prognostic factors):			
Number and type of prognostic factors (eg, obtained from demographics, patient history, physical examination, additional testing, disease characteristics)	X		X
Definition and method for measurement of prognostic factors		X	X
Timing of prognostic factor measurement (eg, at patient presentation, diagnosis, treatment initiation, at the 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 analysis (eg, 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 or events	X		
Number of outcomes or events in relation to the number of candidate prognostic factors (events per variable)			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 (eg, complete case analysis, imputation, or other methods)			X
Analysis:			
Modelling method (eg, linear, logistic, Cox, parametric survival, competing risks) regression)	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 (eg, all candidate prognostic factors considered, preselection of established prognostic factors, retain only those significant from univariable analysis)			X
Method for selection or exclusion of prognostic factors (including those of interest and those used as adjustment factors) during multivariable modelling (eg, backward or forward selection, or full model approach including all factors regardless), and criteria used for any selection or exclusion (eg, P value, Akaike information criterion)			X
Method of handling each continuous prognostic factor (eg, dichotomisation, categorisation, linear, non-linear), including values of any cutpoints used and their justification; for non-linear trends, the method of identifying non-linear relationships (eg, splines, fractional polynomials)			X
Results:			
Unadjusted and adjusted prognostic effect estimates (eg, 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	X
For each extracted adjusted prognostic effect estimate of interest, the set of adjustment factors used	X	X	X
Interpretation and discussion:			
Interpretation of presented results	X	X	
Comparison with other studies, discussion of generalisability, strengths and limitations	X	X	

CHARMS=checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies. CHARMS-PF enables reviewers to describe, assess (eg, for applicability or risk of bias), and summarise (individually and within a meta-analysis) primary studies.

For example, consider a logistic regression of a binary outcome including three adjustment factors (A_1, A_2, and A_3) and one new index prognostic factor (X_1), which is expressed as:

$$\ln(p/(1-p)) = \alpha+\beta_1A_1+\beta_2A_2+\beta_3A_3+\beta_4X_1$$

Here, "p" is the probability of the outcome. After estimation of all the unknown parameters (that is, α, β_1, β_2, β_3, β_4), of key interest is the estimated β_4. This parameter provides the adjusted prognostic effect of the index prognostic factor and reveals its independent contribution to the prediction of the outcome over and above the prognostic effects of the other (established comparator) 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, in which 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

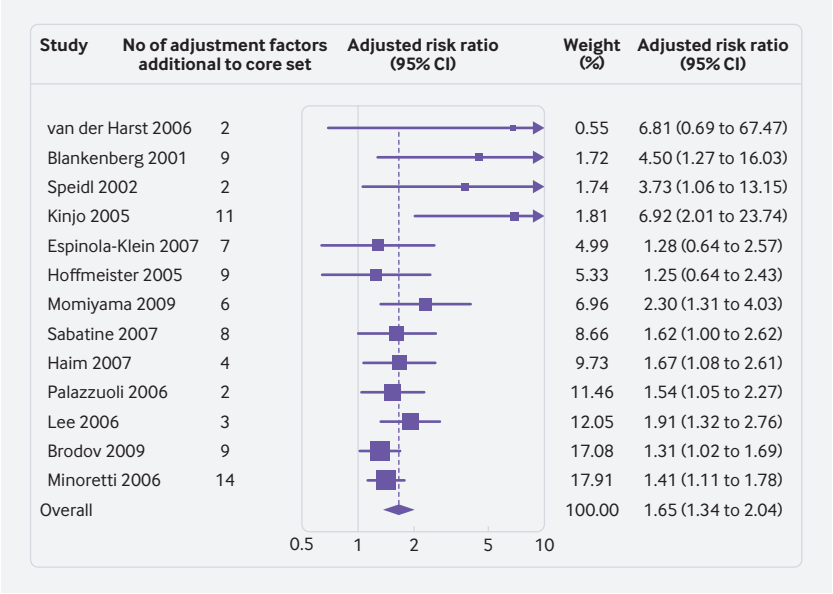

Fig 1 | Forest plot showing the study specific estimates and meta-analysis summary result of the adjusted prognostic effect (risk ratio) of C reactive protein taken from the review of Hemingway and colleagues[14]; all studies were adjusted for a core set of existing prognostic factors (age, sex, smoking status, diabetes, obesity, and lipids), plus up to 14 other prognostic factors. 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 and colleagues,[14] but the confidence interval is wider because we used the Hartung-Knapp approach to account for uncertainty in variance estimates.[16] Although "risk ratio" is used, the estimates actually correspond to a mixture of risk ratios, odds ratios, and hazard ratios

factors," which is the term typically used for prognosis research. The crude (unadjusted) prognostic effect of some index factors may completely disappear after adjustment and is therefore rather uninformative, especially because prognostication in healthcare is rarely based on a single prognostic factor but rather on the information from multiple prognostic factors.[4]

This article focuses on systematic reviews to summarise prognostic factor effect estimates. Some primary studies may also evaluate an index factor's added value in terms of improvement in risk classification and clinical use (eg, measures such as net reclassification improvement and net benefit), and change in prediction model performance (eg, by calculating the change in the concordance index, also known as the C statistic or area under the receiver operating characteristics (ROC) curve).[17-20] However, this is beyond the scope of this article, and we refer the reader to other relevant sources.[6 21 22]

Application to CRP review

CRP is widely studied for its prognostic value in patients with coronary disease. However, there is uncertainty whether CRP is useful because US and European clinical practice guidelines recommend measurement but clinical practice varies widely. This uncertainty motivated the systematic review by Hemingway and colleagues,[14] with the corresponding PICOTS system presented in box 1. No studies were excluded on the basis of methodological standards, sample size, duration of follow-up, publication year, or language of publication.

Box 1: Six items (PICOTS) defining the question for systematic reviews of prognostic factor studies, based on CHARMS (checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies)[6 15] and applied to a review of the adjusted prognostic value of C reactive protein (CRP)[14]

- Population: define the target population for which prognostic factors under review are to be used. For example, CRP review: patients with stable coronary disease, defined as clinically diagnosed angina pectoris or angiographic disease, or a history of acute coronary syndrome at least two weeks before prognostic factor (CRP) measurement.
- Index prognostic factor: define the factors for which prognostic value is under review. For example, CRP review: CRP was the single biomarker reviewed for its prognostic value.
- Comparator prognostic factors: comparator prognostic factors can be considered in a review in various ways. For example, the aim could be to compare the prognostic ability of a certain index factor with two or more other (that is, comparator) prognostic factors; or to review the adjusted prognostic value of a particular index factor—that is, over and above (adjusted for, independent of) other existing (that is, comparator) prognostic factors. If the only aim is to summarise the unadjusted prognostic effect of a particular index factor, then no comparator factor is being considered. For example, CRP review: the focus was on the adjusted prognostic value of CRP—that is, 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, low density lipoprotein cholesterol, high density lipoprotein cholesterol, triglycerides) and inflammatory markers (fibrinogen, interleukin 6, white cell count).
- Outcome: define the outcomes for which the prognostic ability of the factor(s) under review are of interest. For example, CRP review: outcome events were defined as coronary (coronary death, sudden cardiac death, acute non-fatal myocardial infarction, primary percutaneous coronary intervention, unplanned emergency admissions with unstable angina), cardiovascular (when coronary events were reported in combination with heart failure, stroke, or peripheral arterial disease), and all cause mortality.
- Timing: define firstly at what time points the prognostic factors (index and comparators) are to be used (that is, the time point of prognostication), and secondly over what time period the outcomes are predicted by these factors. For example, CRP review: 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 setting and role of the prognostic factors under review. For example, CRP review: CRP measurement was studied in primary and secondary care to provide prognostic information about patients diagnosed with coronary heart disease; this information may be useful for healthcare professionals treating and managing such patients.

Step 2: Searching for and selection of eligible studies

The next step is to identify primary studies that are eligible for 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 with studies of interventions, there is much more variation in the design of prognostic factor studies (eg, data from cohort studies, randomised trials, routine care registries, and case-control studies can all be used), patient inclusion criteria, prognostic factor and outcome measurement, follow-up time, methods of statistical analysis, and adjustment of (and number of) other prognostic factors (covariates). Between-study heterogeneity is therefore the rule rather than the exception in prognostic factor research. It is essential that systematic reviews of prognostic factor studies define the study inclusion and exclusion criteria based on the PICOTS structure (step 1) because this determines the study search and selection strategy.

Typically, broad search and selection filters are required that combine 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).[23] A broad search comes at the (often considerable) expense of retrieving many irrelevant records. Geersing and colleagues[24] validated various existing search strategies for prognosis studies and suggested a generic filter for identifying studies of prognostic factors,[23 25 26] which extended the work of Ingui, Haynes, and Wong.[23 25 26] When tested in a single review of prognostic factors, this generic filter had a number needed to read of 569 to identify one relevant article, emphasising the difficulty in targeting prognostic factor articles. The number needed to read could be considerably reduced when specific factors or populations are added to the filter. Even then, care is needed to be inclusive because multiple terms are often used for the same meaning; for example, biomarker MYCN is also referred to as n-myc and nmyc, among other terms.[13]

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

Application to CRP review

Hemingway and colleagues 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 before CRP measurement. Hemingway and colleagues 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 1566 articles, of which 83 fulfilled the inclusion criteria. If specific terms for CRP had not been included in the search string, then the total number of identified articles would have far exceeded 1566.

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 to the review question and their risk of bias (see step 4). This step also provides the information required for subsequent qualitative and quantitative (meta-analysis) synthesis of the evidence across studies. The CHARMS checklist gives explicit guidance (table 2 in the article by Moons and colleagues[15]) about which key items across 11 domains should be extracted from primary studies of prediction models, and for what reason (that is, to provide general information about the primary study, to guide risk of bias assessment, or to assess applicability of the primary study to the review question). Based on our 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 in reviews of prognostic factors (here referred to as CHARMS-PF; table 1). This basically means that three domains typically addressing multivariable prediction modelling aspects 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 are similar to CHARMS for prediction models. Because we developed the original CHARMS checklist, a wider consensus of the CHARMS-PF content was not considered necessary.

Reviewers should extract fundamental information from the primary prognostic factor 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 in this information across studies. The extracted information can be summarised in tables of study characteristics. In addition, more specific information is needed to properly assess applicability and risk of bias (see step 4), such as methods used to measure prognostic factors and outcomes, handling missing data, attrition (loss to follow-up), and whether estimated associations of the

Table 2 | QUIPS tool (quality in prognostic factor studies), which can be used to classify risk of bias of prognostic factor studies

Domains	Signalling items	Risk of bias ratings
1. Study participation	(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	High: the relationship between the PF and outcome is very likely to be different for participants and eligible non-participants Moderate: the relationship between the PF and outcome may be different for participants and eligible non-participants Low: the relationship between the PF and outcome is unlikely to be different for participants and eligible non-participants
2. Study attrition	(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	High: the relationship between the PF and outcome is very likely to be different for completing and non-completing participants Moderate: the relationship between the PF and outcome may be different for completing and non-completing participants Low: the relationship between the PF and outcome is unlikely to be different for completing and non-completing participants
3. Prognostic factor measurement	(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 cutpoints 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	High: the measurement of the PF is very likely to be different for different levels of the outcome of interest Moderate: the measurement of the PF may be different for different levels of the outcome of interest Low: the measurement of the PF is unlikely to be different for different levels of the outcome of interest
4. Outcome measurement	(a) A clear definition of the outcome is provided (b) Method of outcome measurement used is adequately valid and reliable (c) The method and setting of outcome measurement is the same for all study participants	High: the measurement of the outcome is very likely to be different related to the baseline level of the PF Moderate: the measurement of the outcome may be different related to the baseline level of the PF Low: the measurement of the outcome is unlikely to be different related to the baseline level of the PF
5. Adjustment for other prognostic factors	(a) All other important PFs are measured (b) Clear definitions of the important PFs measured are provided (c) Measurement of all important PFs is adequately valid and reliable (d) The method and setting of PF measurement are the same for all study participants (e) Appropriate methods are used to deal with missing values of PFs, such as multiple imputation (f) Important PFs are accounted for in the study design (g) Important PFs are accounted for in the analysis	High: the observed effect of the PF on the outcome is very likely to be distorted by another factor related to PF and outcome Moderate: the observed effect of the PF on outcome may be distorted by another factor related to PF and outcome Low: the observed effect of the PF on outcome is unlikely to be distorted by another factor related to PF and outcome
6. Statistical analysis and reporting	(a) Sufficient presentation of data to assess the adequacy of the analytic strategy (b) Strategy for model building is appropriate and is based on a conceptual framework or model (c) The selected statistical model is adequate for the design of the study (d) There is no selective reporting of results	High: the reported results are very likely to be spurious or biased related to analysis or reporting Moderate: the reported results may be spurious or biased related to analysis or reporting Low: the reported results are unlikely to be spurious or biased related to analysis or reporting

PF=prognostic factor. Some wording from Hayden and colleagues[27] has been modified to be consistent with the terminology used in this article.

prognostic factors under review were adjusted for other prognostic factors. This information 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), 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 (that is, non-constant hazard ratios for the prognostic factors of interest) should also be extracted; however, such information is rarely reported in sufficient detail.

Unfortunately, many prognostic factor studies do not adequately report estimated prognostic effect measures or their precision. For this reason, methods are available to restore the missing information upon data extraction. In particular, Parmar and colleagues[28] and Tierney and colleagues[29] 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 (eg, 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 (eg, "high" versus "normal" levels). Perneger and colleagues[30] report how to derive unadjusted hazard ratios from survival proportions, and Pérez and colleagues suggest using a simulation approach.[31] Even with such indirect estimation methods, not all results can be obtained. For example, in a systematic review of 575 studies investigating prognostic factors in neuroblastoma,[32] the methods of Parmar and colleagues were used to obtain 204 hazard ratio estimates and their confidence intervals; but this represented only 35.5% of the potential evidence.

Although indirect estimation methods help retrieve unadjusted prognostic factor effect estimates, they often have limited value for obtaining adjusted effect

estimates. Furthermore, even when multiple studies provide the adjusted prognostic effect of a particular factor, the set of adjustment factors will usually differ across studies, which complicates the interpretation of subsequent meta-analysis results. We recommend that reviewers predefine the core set of prognostic factors for the outcome of interest (eg, age, sex, smoking status, disease stage) that represents the desired "minimal" set of adjustment factors. An agreed process among health professionals and researchers in the field could be required to define 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 the prognostic 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" versus "high" group. When the outcome is defined differently across studies, approaches to convert effect measures on different outcome scales could be useful.[33] Also, to deal with different cutpoint levels for values of a particular factor,[34] the prognostic effects of "high" versus "normal" could be converted to prognostic effects relating to a 1 unit increase in the factor. This requires assumptions about the underlying distribution of the factor. Such an approach was used by Hemingway and colleagues.[14] Of 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 standardised effect estimates by standardising the corresponding regression coefficients.[35]

Application to CRP review

Hemingway and colleagues extracted background information such as year of study start, number of included patients, mean age, baseline coronary morbidity (eg, 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 the reported risk ratio, odds ratio, or hazard ratio (labelled generally as "risk ratio" in this article), and 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. If 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 are identified and data are extracted, an important next step is to assess the applicability and risk of bias (quality) of each study in the review. As for steps 2 and 3, ideally this is done by two reviewers, independently, with any discrepancies resolved. 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. Some aspects of a study may be applicable (eg, correct condition at start point, with prognostic factors of interest evaluated) but not others (eg, incorrect population or setting, inappropriate outcome definition, different follow-up time, lack of adjustment for conventional prognostic factors). Applicability is typically first assessed during title and abstract screening, and then during this step, so that it is based on full text screening and determined by PICOTS (step 1) and inclusion and exclusion criteria of studies (step 2).

Risk of bias refers to the extent to which flaws in the study design or analysis methods could 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.[8 32 36-44] For prognostic factor studies, Hayden and colleagues developed the QUIPS checklist (quality in prognostic factor studies) for examining risk of bias across six domains[27]: study participation, study attrition, prognostic factor measurement, outcome measurement, adjustment for other prognostic factors, and statistical analysis and reporting. Table 2 shows the signalling items within these domains to help guide reviewers in making low, unclear, or high risk of bias classifications. Additional guidance may be found in general tools examining the quality of observational studies,[45 46] and the REMARK guideline (reporting recommendations for tumour marker prognostic studies) for reporting of primary prognostic factor studies.[47 48]

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 could 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 factors and outcomes should also be specified at the protocol stage. Similarly, the core set of other (adjustment) prognostic factors that are deemed necessary for the primary studies to have adjusted for, should be predefined to facilitate judgment related to risk of bias in 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 needed to assess the bias domains is not available in the study report. When this occurs, other publications that may have used the same dataset (which often occurs in prognostic studies based on large existing cohorts) should be consulted and study authors should be contacted for additional information. An informed judgment 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 separately because this will inform readers, flag improvements needed for subsequent primary studies, and facilitate future meta-epidemiological research. We recommend defining studies with an overall "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 to moderate) risk of bias.

Application to CRP review

Hemingway and colleagues assessed the quality of included studies by the quality of their reporting on 17 items derived from the REMARK guideline.[48] The median number of study quality items reported was seven 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 and colleagues noted 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 and this raised applicability concerns.

Step 5: Meta-analysis

Meta-analysis of prognostic factor studies aims to summarise the (adjusted) prognostic effect of each factor of interest. In addition to missing estimates, challenges for the meta-analyst include (a) having different types of prognostic effect measures (eg, odds ratios and hazard ratios), which are not necessarily comparable[30]; (b) estimates without standard errors, which is a problem because meta-analysis methods typically weight each study by (a function of) their standard error; (c) estimates relating to various time points of the outcome occurrence or measurement; (d) different methods of measurement for prognostic factors and outcomes; (e) various sets of adjustment factors; and (f) different approaches to handling continuous prognostic factors (eg, categorisation, linear, non-linear trends), including the choice of cutpoint value when dichotomising continuous values into "high" and "normal" groups. Many of these issues lead to substantial heterogeneity and if a meta-analysis is performed, summary results cannot be directly interpreted.

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

- Hazard ratios, odds ratios, and risk ratios separately
- Unadjusted and adjusted associations separately
- Prognostic factor effects at distinct cutpoints (or groups of similar cutpoints) 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.

Ideally a meta-analysis of adjusted results should ensure that all included estimates are adjusted for the same set of other prognostic factors. This situation 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 (predefined) minimum set of adjustment factors (that is, a core set of established prognostic factors).

Even when adhering to this guidance, unexplained heterogeneity is likely to remain because of other reasons (eg, differences in length of follow-up or in treatments received during follow-up). Therefore, if a meta-analysis is performed, a random effects approach is essential to allow for unexplained heterogeneity across studies (box 2), as previously described in *The BMJ*.[53] This approach provides a summary estimate of the average prognostic effect of the index factor and the variability in effect across studies. Also potentially useful are meta-analysis methods to estimate the trend (eg, linear effect) of a prognostic factor that has been grouped into three or more categories within studies (with each category compared with the reference category). These methods generally model the estimated prognostic effect sizes in each category as a function of "exposure" level (eg, midpoint or median prognostic factor value in the category) and account for within-study correlation and between-study heterogeneity.[54-58] To apply these methods, some additional knowledge of the factor's underlying distribution is usually needed to help define the "exposure" level because the chosen value can have an impact on the results.[56]

Application to CRP review

Hemingway and colleagues[14] applied 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, sex, smoking status, diabetes, obesity, and lipids). The summary meta-analysis result was a risk ratio of 1.97 (95% confidence interval 1.78 to 2.17), which gives the average prognostic effect of CRP (for those in the top *v* 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 (that is, risk ratio >1). When restricting meta-analysis to just the

Box 2: 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; therefore assuming a common (fixed) prognostic effect is not sensible. If Y_i and var(Y_i) denote the prognostic effect estimate (eg, ln(hazard ratio), ln(odds ratio), ln(risk ratio), or mean difference) and its variance in study *i*, then a general random effects meta-analysis model can be specified as:

$Y_i \sim N(\mu, var(Y_i)+\tau^2)$.

Most researchers use either restricted maximum likelihood or the approach of DerSimonian and Laird to estimate this model,[49] but other options are available, including a Bayesian approach.[50] Of key interest is the estimate of μ, which reveals the summary (average) prognostic effect of the index prognostic factor of interest. The standard deviation of this prognostic factor 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),[51] and we have found the approach of Hartung-Knapp to be robust for this purpose in most settings.[16 52] 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.

Advanced multivariate meta-analysis methods are also available to handle multiple cutpoints,[59] multiple methods of measurement,[59] or different adjustment factors in prognostic factor studies.[60] An introduction to multivariate meta-analysis has been published in *The BMJ*.[61]

13 studies that adjusted for at least all six conventional prognostic factors, the summary risk ratio decreased to 1.65 (95% confidence interval 1.39 to 1.96), and the between-study heterogeneity reduced. Using the study specific estimates given by Hemingway and colleagues, we updated this meta-analysis (fig 1), obtaining the same summary result but a wider confidence interval (1.34 to 2.04) through the Hartung-Knapp approach.[16]

Step 6: Quantifying and examining heterogeneity

For all meta-analyses, when there is large heterogeneity across included studies, it may be better not to synthesise the study results, but rather display the variability in estimates on a forest plot without showing an overall pooled estimate. When a meta-analysis is performed in the face of heterogeneity, it is important to quantify and report the magnitude of heterogeneity itself; for example, through the estimate of (the between-study variance),[62] or an approximate 95% prediction interval indicating the potential true prognostic effect of a factor in a new population.[53 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 a low risk of bias, those with a follow-up of less than one year or of at least one year, or those set in countries in Europe. A better approach is meta-regression, which extends the meta-analysis equation shown in box 2 by including study level covariates,[64] and allows a formal comparison of meta-analysis results across groups defined by covariates (eg, low risk of bias studies *v* 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 so 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 cutpoint level compared with studies at higher risk of bias.

Application to CRP review

Hemingway and colleagues reported 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 factors, the (log) number of events, and the proportion of patients with stable coronary disease (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 risk ratio decreased by 3%. The summary risk ratio 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, or by the type of prognostic effect measure provided (that is, risk ratio, odds ratio, or hazard ratio).

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 because these often evaluate many potential prognostic factors with relatively few outcome events) show larger prognostic effects than larger studies. This difference may be due to chance or heterogeneity, but a major threat here is publication bias and selective reporting, which are endemic in prognosis research.[36-38] Such reporting biases lead to smaller studies, with (statistically) significant or larger prognostic factor effect estimates being more likely to be published or reported in sufficient detail, and thus included in a meta-analysis, than smaller studies with non-significant or smaller prognostic effect estimates. This bias is a potential concern for unadjusted and adjusted prognostic effects. A primary study usually estimates an unadjusted prognostic effect for each of multiple prognostic factors, but study authors may only report effects that are statistically significant. In addition, adjusted results are often only reported for prognostic factors that retain statistical significance in univariable and multivariable analysis. A consequence is that meta-analysis results will be biased, with larger summary prognostic effects than in reality, and potentially some factors being deemed to have clinical value when actually they 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). A funnel plot is usually recommended if there are 10

or more studies.[65] The plot should ideally show a symmetric, funnel like shape, with results from larger studies at the centre of the funnel and 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. An example is shown in figure 2.

As mentioned, small-study effects may also arise due to heterogeneity. Therefore, 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 it being caused by publication bias. A multivariate meta-analysis could reduce the impact of small-study effects by "borrowing strength" from related information.[61]

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 design leads to spurious (due to chance) and potentially biased (due to poor estimation properties[68]) prognostic effect estimates, which are more prone to selective reporting. 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] Therefore, larger studies are less likely to identify spurious prognostic factor effect estimates. It is helpful to examine small-study effects (potential publication bias) when restricting analysis to the subset of studies at low risk of bias. If this approach 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 only on the higher quality studies.

Application to CRP review

Figure 2 shows a funnel plot of the study estimates from the CRP meta-analysis shown in figure 1. There is clear asymmetry, which shows the strong potential for 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.

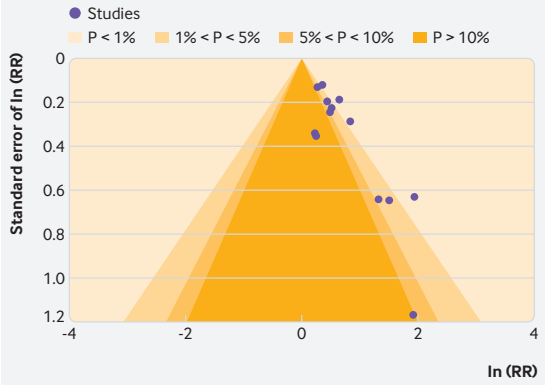

Fig 2 | Evidence of funnel plot asymmetry (small-study effects) in the C reactive protein meta-analysis shown in figure 1. The smaller studies (with higher standard errors) have risk ratio (RR) estimates mainly to the right of the larger studies, and therefore give the largest prognostic effect estimates. A concern is that this is due to publication bias, with "missing" studies potentially falling to the left side of the larger studies and in the lighter shaded regions 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 (preferred reporting items for systemic reviews and meta-analyses) and MOOSE (meta-analysis of observational studies in epidemiology) will be relevant,[69 70] and should be complemented by REMARK,[47 48] 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 is the essential input for this step. Discussion is necessary on whether and how the prognostic factors identified may be useful in practice (that is, translation of results to clinical practice), and what further research is necessary. Ideally impact studies (eg, randomised trials that compare groups which do and do not use a prognostic factor to inform clinical practice) are needed before strong recommendations for clinical practice are made; however, these studies are rare and outside the scope of the review framework outlined in this article.

To interpret the certainty (confidence) of the summary results of a review of intervention effectiveness, GRADE (grades of recommendation, assessment, development, and evaluation) was developed. This approach assesses the overall quality of and certainty in 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 can be assessed using the information obtained by the tools and methods described in the above steps. However, it is not known whether these domains, developed for reviews of interventions, are equally applicable to assessing the certainty of summary results of systematic reviews of prognostic factor studies. Compared with reviews of intervention studies, allowing for heterogeneity (the inconsistency domain) might be more acceptable in reviews of

doi: 10.1136/bmj.k4597 | BMJ 2019;364:k4597 | the bmj

prognostic factor studies because of the inevitable heterogeneity caused by study differences in methods of measurement, adjustment factors, and statistical analysis methods, among others. Furthermore, the threat of selective reporting or publication bias in reviews of prognostic factor studies may be more severe than in reviews of intervention studies because of the problems of exploratory studies, poor reporting, and biased analysis methods.

There is limited empirical evidence for using the existing domains to grade the certainty of summary estimates of prognostic factor studies, although a first attempt has been made[71]; in addition, an assessment has been performed on grading the certainty of evidence of summary estimates of overall prognosis studies.[72] Reviewers need to be especially cautious when comparing the adjusted prognostic value of multiple index factors, for example, to conclude whether the summary adjusted hazard ratio for 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] Even then, due to different scales and distributions of each factor (eg, continuous or binary), a simple comparison of the prognostic effect sizes (eg, hazard ratio for factor A *v* hazard ratio for factor B) may not be straightforward.

Application to CRP review

The meta-analysis results suggest CRP is a prognostic factor for the risk of death and non-fatal cardiovascular events, even when only including the largest studies that adjusted for all six conventional prognostic factors. In their discussion, Hemingway and colleagues downgraded the meta-analysis findings because of a strong concern about the quality and reliability of the underlying evidence.[14] The absence of prespecified protocols, poor and potentially biased reporting, and strong potential for publication bias prevented the authors from making firm conclusions about whether CRP has prognostic value 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

In this article, 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 76] We expect the prevalence of such reviews to grow rapidly, especially as Cochrane has recently embarked on prognosis reviews (see also the Cochrane Prognosis Methods Group website www.methods.cochrane.org/prognosis).[77] Our guidance will help researchers to write grant applications for reviews of prognostic factor studies, and to develop protocols and conduct such reviews. Protocols of prognostic factor reviews 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.[77] Our guidance will also allow readers and healthcare providers to better judge reports of prognostic factor reviews.

Finally, we note that some of the limitations described (eg, use of different cutpoint values across studies) could be alleviated if the individual participant data were obtained from primary prognostic factor studies[78] rather than being extracted from study publications; although, this may not solve all problems (eg, quality of original study, availability of different adjustment factors).[79] Further discussion on individual participant data meta-analysis of prognostic factor studies is given elsewhere.[80]

We thank the editors of *The BMJ* and the three reviewers for their helpful feedback that improved the article on revision.

Contributors: RDR and KGMM contributed equally, conceived the article content and structure, and wrote the first draft. RDR, KIES, JE, and TD added application to the C reactive protein review. All authors provided intellectual content, text, and corrections to improve the first draft. KGMM, DGA, and GSC codeveloped the CHARMS guidance that informed the content of this paper. JH codeveloped the QUIPS checklist, and informed the use and interpretation of QUIPS for this paper. RDR, KGMM, and JH and subsequently all other authors revised the article after comments received by reviewers and *The BMJ*. RDR is the guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: RDR and KIES 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; project No 390). KIES is also supported by a launching fellowship from the NIHR SPCR. TPAD 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, NHS, Department of Health, or the Netherlands Organisation for Health Research and Development. GSC was supported by the NIHR Biomedical Research Centre, Oxford. We gratefully acknowledge the Cochrane Methods Innovation Fund and the Cochrane Strategic Methods Fund for their contribution.

Competing interests: None.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Leeflang MMG, Deeks JJ, Gatsonis C, Bossuyt PM, Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008;149:889-97. doi:10.7326/0003-4819-149-12-200812160-00008
2 Hemingway H, Croft P, Perel P, et al, PROGRESS Group. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ* 2013;346:e5595. doi:10.1136/bmj.e5595
3 Riley RD, Hayden JA, Steyerberg EW, et al, PROGRESS Group. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med* 2013;10:e1001380. doi:10.1371/journal.pmed.1001380
4 Steyerberg EW, Moons KG, van der Windt DA, et al, PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10:e1001381. doi:10.1371/journal.pmed.1001381

5 Hingorani AD, Windt DA, Riley RD, et al, PROGRESS Group. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ* 2013;346:e5793. doi:10.1136/bmj.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. doi:10.1136/bmj.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. doi:doi:10.1177/0962280218785504
8 Bouwmeester W, Zuithoff NP, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9:1-12. doi:10.1371/journal.pmed.1001221
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:W1-73. doi:10.7326/M14-0698
10 Altman DG. Systematic reviews of evaluations of prognostic variables. *BMJ* 2001;323:224-8. doi:10.1136/bmj.323.7306.224
11 Altman DG, Riley RD. Primer: an evidence-based approach to prognostic markers. *Nat Clin Pract Oncol* 2005;2:466-72. doi:10.1038/ncponc0287
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;100:1219-29. doi:10.1038/sj.bjc.6604999
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;7:1-162. doi:10.3310/hta7050
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:e1000286. doi:10.1371/journal.pmed.1000286
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:e1001744. doi:10.1371/journal.pmed.1001744
16 Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med* 2001;20:3875-89. doi:10.1002/sim.1009
17 Royston P. Explained variation for survival models. *Stata J* 2006;6:83-96.
18 Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;23:723-48. doi:10.1002/sim.1621
19 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. doi:10.1136/bmj.i6
20 Pencina MJ, D'Agostino RBSr, D'Agostino RBJr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157-72, discussion 207-12. doi:10.1002/sim.2929
21 Pennells L, Kaptoge S, White IR, Thompson SG, Wood AM, Emerging Risk Factors Collaboration. Assessing risk prediction models using individual participant data from multiple studies. *Am J Epidemiol* 2014;179:621-32. doi:10.1093/aje/kwt298
22 Wynants L, Riley RD, Timmerman D, Van Calster B. Random-effects meta-analysis of the clinical utility of tests and prediction models. *Stat Med* 2018;37:2034-52. doi:10.1002/sim.7653
23 Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc* 2001;8:391-7. doi:10.1136/jamia.2001.0080391
24 Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews [correction in: *PLoS One* 2012;7: doi/10.1371/annotation/96bdb520-d704-45f0-a143-43a48552952e]. *PLoS One* 2012;7:e32844. doi:10.1371/journal.pone.0032844
25 Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR, Hedges Team. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ* 2005;330:1179. doi:10.1136/bmj.38446.498542.8F
26 Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R, Hedges Team. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. *AMIA Annu Symp Proc* 2003;728-32.
27 Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. *Ann Intern Med* 2013;158:280-6. doi:10.7326/0003-4819-158-4-201302190-00009
28 Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat Med* 1998;17:2815-34. doi:10.1002/(SICI)1097-0258(19981230)17:24<2815::AID-SIM110>3.0.CO;2-8
29 Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007;8:16. doi:10.1186/1745-6215-8-16
30 Perneger TV. Estimating the relative hazard by the ratio of logarithms of event-free proportions. *Contemp Clin Trials* 2008;29:762-6. doi:10.1016/j.cct.2008.06.002
31 Pérez 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. doi:10.1016/j.jclinepi.2017.12.017
32 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;88:1191-8. doi:10.1038/sj.bjc.6600886
33 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.* John Wiley & Sons, 2009:45-49. doi:10.1002/9780470743386.ch7.
34 Sadashima E, Hattori S, Takahashi K. Meta-analysis of prognostic studies for a biomarker with a study-specific cutoff value. *Res Synth Methods* 2016;7:402-19. doi:10.1002/jrsm.1201
35 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. *Epidemiol Biostatistics Public Health* 2013;10:e8854.
36 Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. *Eur J Cancer* 2007;43:2559-79. doi:10.1016/j.ejca.2007.08.030
37 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;99:236-43. doi:10.1093/jnci/djk032
38 Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. *J Natl Cancer Inst* 2005;97:1043-55. doi:10.1093/jnci/dji184
39 Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Med* 2010;8:20. doi:10.1186/1741-7015-8-20
40 Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med* 2011;9:103. doi:10.1186/1741-7015-9-103
41 Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013;66:268-77. doi:10.1016/j.jclinepi.2012.06.020
42 Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *Br J Cancer* 2004;91:4-8. doi:10.1038/sj.bjc.6601907
43 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. doi:10.1186/1471-2288-14-40
44 Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. *BMC Med* 2010;8:21. doi:10.1186/1741-7015-8-21
45 Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919. doi:10.1136/bmj.i4919
46 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.
47 Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. *PLoS Med* 2012;9:e1001216. doi:10.1371/journal.pmed.1001216
48 McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. REporting recommendations for tumour MARKer prognostic studies (REMARK). *Br J Cancer* 2005;93:387-91. doi:10.1038/sj.bjc.6602678
49 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177-88. doi:10.1016/0197-2456(86)90046-2
50 Langan D, Higgins JP, Simmonds M. An empirical comparison of heterogeneity variance estimators in 12 894 meta-analyses. *Res Synth Methods* 2015;6:195-205. doi:10.1002/jrsm.1140
51 Cornell JE, Mulrow CD, Localio R, et al. Random-effects meta-analysis of inconsistent effects: a time for change. *Ann Intern Med* 2014;160:267-70. doi:10.7326/M13-2886
52 Partlett C, Riley RD. Random effects meta-analysis: Coverage performance of 95% confidence and prediction intervals following REML estimation. *Stat Med* 2017;36:301-17. doi:10.1002/sim.7140

53 Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011;342:d549. doi:10.1136/bmj.d549
54 Berlin JA, Longnecker MP, Greenland S. Meta-analysis of epidemiologic dose-response data. *Epidemiology* 1993;4:218-28. doi:10.1097/00001648-199305000-00005
55 Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *Am J Epidemiol* 1992;135:1301-9. doi:10.1093/oxfordjournals.aje.a116237
56 Hartemink N, Boshuizen HC, Nagelkerke NJ, Jacobs MA, van Houwelingen HC. 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;163:1042-52. doi:10.1093/aje/kwj141
57 Shi JQ, Copas JB. Meta-analysis for trend estimation. *Stat Med* 2004;23:3-19, discussion 159-62. doi:10.1002/sim.1595
58 Orsini N, Li R, Wolk A, Khudyakov P, Spiegelman D. Meta-analysis for linear and nonlinear dose-response relations: examples, an evaluation of approximations, and software. *Am J Epidemiol* 2012;175:66-73. doi:10.1093/aje/kwr265
59 Riley RD, Elia EG, Malin G, Hemming K, Price MP. Multivariate meta-analysis of prognostic factor studies with multiple cut-points and/or methods of measurement. *Stat Med* 2015;34:2481-96. doi:10.1002/sim.6493
60 Jackson D, White I, Kostis JB, et al, Fibrinogen Studies Collaboration. Systematically missing confounders in individual participant data meta-analysis of observational cohort studies. *Stat Med* 2009;28:1218-37. doi:10.1002/sim.3540
61 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. doi:10.1136/bmj.j3932
62 Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Med Res Methodol* 2008;8:79. doi:10.1186/1471-2288-8-79
63 Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009;172:137-59. doi:10.1111/j.1467-985X.2008.00552.x
64 Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Stat Med* 1995;14:395-411. doi:10.1002/sim.4780140406
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;343:d4002. doi:10.1136/bmj.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;9:41-50. doi:10.1002/jrsm.1266
67 Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA* 2006;295:676-80. doi:10.1001/jama.295.6.676
68 Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993;80:27-38. doi:10.1093/biomet/80.1.27 .
69 Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535. doi:10.1136/bmj.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;283:2008-12. doi:10.1001/jama.283.15.2008
71 Huguet A, Hayden JA, Stinson J, et al. Judging the quality of evidence in reviews of prognostic factor research: adapting the GRADE framework. *Syst Rev* 2013;2:71. doi:10.1186/2046-4053-2-71
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. doi:10.1136/bmj.h870
73 Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Ann Intern Med* 2013;158:544-54. doi:10.7326/0003-4819-158-7-201304020-00006
74 Pearson TA, Mensah GA, Alexander RW, et al, Centers for Disease Control and Prevention, American Heart Association. 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;107:499-511. doi:10.1161/01.CIR.0000052939.59093.45
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. *Commun Stat* 2006;35:1333-42. doi:10.1080/03610920600629666 .
76 Peat G, Riley RD, Croft P, et al, PROGRESS Group. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. *PLoS Med* 2014;11:e1001671. doi:10.1371/journal.pmed.1001671
77 Moons KG, Hooft L, Williams K, Hayden JA, Damen JA, Riley RD. Implementing systematic reviews of prognosis studies in Cochrane. *Cochrane Database Syst Rev* 2018;10:ED000129.
78 Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010;340:c221. doi:10.1136/bmj.c221
79 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.* Birkhäuser, 2006: 3-18.
80 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. doi:10.1186/1471-2288-12-56