### Supplement: Statistical analysis plan

### Measures of association or predictive performance measures to be extracted

We extracted information about the models’ predictive performance, in terms of discrimination (C-statistic) and calibration (calibration slope, ratio of observed (O) to expected (E) events (O:E ratio), calibration plots), and net benefit measures.

### Dealing with missing data

When performance measures (such as C-statistic, O:E ratio) were not reported in the paper, we contacted authors. Where possible, we used standard methods and formulae described by Debray and colleagues to estimate the O:E ratio and C-statistic and associated standard errors ([Debray 2017](#REF-Debray-2017)).

### Assessment of heterogeneity

Reviews of prognostic studies often have to deal with a substantial amount of heterogeneity. We planned to assess the impact of heterogeneity in predictive performance across validation studies, where there were enough data to do so, by calculating prediction intervals that provide a range for the potential performance of a model in a new validation study ([Debray 2017](#REF-Debray-2017)). We also planned to calculate I2 and Tau2 statistics. If reported, we would have extracted performance in subgroups.

## Data synthesis

### Data synthesis and meta-analysis approaches

If there were enough studies reporting external validation performance, we planned to conduct random-effects meta-analyses to summarise performance of prognostic models, as data were likely to be highly heterogeneous. We aimed to pool information about each model’s discrimination (using C-statistic or equivalent), calibration (using calibration slope, calibration-in-the-large; and O:E ratio) and equivalents from time-to-event models (e.g. Harrell’s C-statistic, calibration slope, D statistic, O:E at each time point). We planned to summarise performance measures separately, first transforming them to an appropriate scale where necessary (logit C-statistic and log O:E ratio) to produce summary results (with 95% confidence intervals (CIs)) that quantified the average performance across studies ([Snell 2018](#REF-Snell-2018)). To better account for the uncertainty in the estimated between-study heterogeneity, we planned to use the restricted maximum likelihood (REML) estimation, with 95% CIs for the summary (average) performance of a model, derived using the Hartung-Knapp-Sidik-Jonkmann method, as recommended by [Debray 2017](#REF-Debray-2017) and [Langan 2018](#REF-Langan-2018). In the absence of sufficient data for a meta-analysis, we have used a narrative synthesis instead.

### Subgroup analysis and investigation of heterogeneity

We planned that, if there were sufficient data (a minimum of 10 studies), we would investigate potential sources of heterogeneity using meta-regression with the summary estimate of model performance (e.g. logit C-statistic or log O:E ratio) as a dependent variable and study-level covariates (population/case-mix (age of participants and multimorbidity), study setting of models (primary and secondary care settings) and study design (follow-up time, source of data, outcome definition and sample size)) as explanatory variables.

### Sensitivity analysis

If we had sufficient studies for meta-analysis, we planned to evaluate the impact of risks of bias by conducting analyses only including studies assessed at low risk of bias.

Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017;356:i6460.

Langan D, Simmonds M. A comparison of heterogeneity variance estimators in simulated random ‐ effects meta ‐ analyses. Research Synthesis Methods 2019;10(1):83-98.

Snell KIE, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: which scale helps ensure between-study normality for the C-statistic and calibration measures? Statistical Methods in Medical Research 2018;27(11):3505–22.