

GRADE concept paper 2: Concepts for Judging Certainty on the Calibration of Prognostic Models in a Body of Validation Studies.

Farid Foroutan, Gordon Guyatt, Marialena Trivella, Nina Kreuzberger, Nicole Skoetz, Richard D. Riley, Pavel S. Roshanov, Ana Carolina Alba, Nigar Sekercioglu, Carlos Canelo, Zachary Munn, Romina Brignardello-Petersen, Holger J. Schünemann, Alfonso Iorio

- (1) Ted Rogers Centre for Heart Research, Peter Munk Cardiac Centre, Toronto, Ontario, Canada
- (2) Department of Health Research Methods, Evidence, and Impact, McMaster University, Ontario, Canada
- (3) Division of Nephrology, Department of Medicine, London Health Sciences Centre, Ontario, Canada
- (4) NK: Cochrane Haematology, Department I of Internal Medicine, Center for Integrated Oncology Aachen Bonn Cologne Duesseldorf, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany
- (5) Evidence-based Oncology, Department I of Internal Medicine, Center for Integrated Oncology Aachen Bonn Cologne Duesseldorf, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany
- (6) School of Medicine, Keele University, Keele, United Kingdom

Address correspondence and requests for reprints to:

Farid Foroutan

Health Research Methods, Evidence, and Impact

McMaster University

foroutaf@mcmaster.ca

Manuscript details

Keywords: GRADE, certainty in evidence, prognosis, prognostic models, systematic review,

Word count:

References:

Tables = 1

Figures = 2

Journal Pre-proof

What is new?

Key findings

- We introduce four concepts underlying rating certainty in calibration of prognostic models in the body of evidence of model validation studies. The first concept focuses on determining the overall inference for which we are rating our certainty (satisfactory vs unsatisfactory model performance). The latter three focus on the application of the GRADE framework to the pooled observed to expected (O:E) risk ratio as the most commonly reported measure of overall calibration in validation studies of prognostic models. The pooled O:E ratio of interest might be that for the whole population, or for particular risk groups or covariate values.

What this adds to what is known?

- The four concepts introduced in this paper provide the necessary steps for applying GRADE to calibration of prediction models (as assessed with the pooled O:E ratio).

What is the implication and what should change?

- When evaluating calibration of prediction models, the extent of inconsistency in the O:E ratio across studies should first inform the overall inference on model performance (satisfactory vs unsatisfactory), and then inform our certainty in the body of evidence.
- O:E ratio provides a suboptimal assessment of calibration, as the O:E ratio may miss important miscalibration, and the pooled O:E may miss important inconsistency of the contributory studies.
- Future validation studies would more effectively report calibration curves, and efficient methods to pool calibration curves are needed.
- Comprehensive GRADE guidance for rating certainty in prognostic models beyond statistical measures of predictive performance and including issues of clinical utility remains under development.

Abstract

Prognostic models combine several prognostic factors to provide an estimate of the likelihood (or risk) of future events in individual patients, conditional on their prognostic factor values. A fundamental part of evaluating prognostic models is undertaking studies to determine whether their predictive performance, such as calibration and discrimination, is reproduced across settings. Systematic reviews and meta-analyses of studies evaluating prognostic models' performance are a necessary step for selection of models for clinical practice and for testing the underlying assumption that their use will improve outcomes, including patient's reassurance and optimal future planning. In this paper, we highlight key concepts in evaluating the certainty of evidence regarding the calibration of prognostic models.

Four concepts are key to evaluating the certainty of evidence on prognostic models' performance regarding calibration. The first concept is that the inference regarding calibration may take one of two forms: deciding whether one is rating certainty that a model's performance is satisfactory or, instead, unsatisfactory, in either case defining the threshold for satisfactory (or unsatisfactory) model performance. Second, inconsistency is the critical GRADE domain to deciding whether we are rating certainty in the model performance being satisfactory or unsatisfactory. Third, depending on whether one is rating certainty in satisfactory or unsatisfactory performance, different patterns of inconsistency of results across studies will inform ratings of certainty of evidence. Fourth, exploring the distribution of point estimates of observed to expected ratio across individual studies, and its determinants, will bear on the need for and direction of future research.

Introduction

This GRADE concept paper presents insights developed by the GRADE prognosis project group in working toward GRADE guidance for rating the certainty in prognostic models. Those already familiar with both concepts of discrimination and calibration in prognostic models, and previous GRADE guidance for interventions and prognosis, may wish to bypass the initial sections of the paper and move immediately to ***The Target of Certainty Rating of Calibration in Systematic Reviews of Prognostic Models.***

Prognostic models and their performance

To help clinicians estimate probability of future health outcomes in their patients, prognostic models simultaneously combine information from multiple prognostic factors (e.g. in a multivariable regression model or a machine learning approach)¹. Such models estimate an individual's risk of a particular outcome occurring in the future (by a particular time-point), and so have the potential to inform clinical decision-making and patient counselling. Clinicians considering use of prognostic models typically rely on external validation studies that test the performance of a model in patients who were not included in the sample on which that model was initially developed². Statistical measures of predictive performance reported in such studies provide a guide for models' potential usefulness in clinical practice.

While some measures are useful in certain situations^{3,4} measures of discrimination and calibration are always important in evaluating the performance of a predictive model. Discrimination provides a measure of how well a model can differentiate (discriminate) between high and low risk patients³, typically relying on quantitative measures that include the c-statistic or area under the receiver operating characteristics curve (AUC),³ which for binary outcomes are equivalent. Calibration measures how well a model's predicted estimates of risk correspond to absolute risks observed in validation datasets, ideally examined across the whole spectrum of predicted risks across individuals⁵.

Due to changes in case-mix, differences in outcome rates, and heterogeneity in prognostic factor effects, the performance of a model may vary on one or both of discrimination and calibration from one external validation study to another⁶. Systematic reviews and meta-analyses of model performance from these validation studies can summarize the best available evidence and help clinicians to draw conclusions regarding the model's predictive performance and thus its potential clinical use⁷. The focus of this paper is rating the certainty in model calibration, while discrimination is not discussed here. Therefore, in the next section we provide a brief review of most common approaches to measure calibration.

Measures of model calibration

A number of different approaches for assessing the calibration of prognostic models exist. Van Calster et al. provide an in-depth review of these approaches and propose a hierarchy of analytic methods for assessing calibration: calibration at large, weak (use of intercept and calibration slope), moderate (use of calibration curve), and strong (plotting calibration in patients with similar patterns of covariates)⁸.

As noted by these authors, the strong approach to assessing calibration often requires an extremely large sample size (potentially even millions of patients). Therefore, primary validation studies seldom report on strong calibration. The best case scenario, however, is validation studies that provide calibration curves (usually corresponding to a particular time-point) with the observed event risk on the y-axis, and the model's predicted risk on the x-axis⁵ (moderate calibration)⁸. Ideally, investigators present a continuous calibration curve, for example using a non-parametric smoother. However, often they will offer categories of patients (e.g. groups defined by tenths of predicted risk). Visual inspection of the calibration curve or plots can address agreement between predicted risk and observed risk. A calibration curve that depicts a straight line ('calibration slope' of 1) at a 45-degree angle, with a y-intercept of 0 suggests a perfectly calibrated model. Calibration curves may depict models that are well-calibrated through particular risk levels (e.g. low or high) but not others^{3 9}. Compared to other measures of calibration, curves are most informative on the degree and patterns of miscalibration. They provide information on performance of model for a wide spectrum of patient risks. For example, calibration curves allow one to discover whether a model significantly under or overestimates risk in low and high-risk individuals.

Lower in the hierarchy of calibration measures is when – instead of a curve – researchers assume a calibration model with a linear relationship⁸. The linear relationship provides an estimate of the intercept and calibration slope. The calibration of the model is assessed with the intercept when the slope is forced to be 1. Most models, however, will show some deviation from perfect calibration, and often also from linearity, so that calibration may be optimal, acceptable or unacceptable for different levels of observed risk^{5 8}. Therefore, reliance on the intercept may mask miscalibration.

Further down in the hierarchy, the observed to expected ratio (O:E) statistic⁷ provides another common measure of calibration (calibration at large)⁸, where $O:E = 1$ suggests that estimated risk is in agreement with observed risk (i.e., good calibration), $O:E > 1$ suggests underestimation, and $O:E < 1.0$ suggests overestimation of

risk. Authors of primary studies often report one O:E measure for the entire cohort, which we designate as an *overall calibration ratio* (averaged across all risk categories within the cohort).

An apparently satisfactory overall calibration ratio may exist despite serious miscalibration. For example, if the model underestimates the risk in half of the patients and overestimates it in the other half, the overall calibration ratio might suggest perfect calibration. To avoid falling in this trap, rather than an average calibration ratio, authors should separately report inferences regarding calibration for different risk categories (e.g., lower or higher risk patients) or for other meaningful subgroup of patients based on characteristics not included in the model (for example, ethnicity, or age beyond that of the derivation cohort).

A similar issue can arise at the study level in meta-analyses of O:E ratio. For example, if half of the studies in one meta-analysis report an O:E <1.0 and the other half >1.0, the pooled average calibration ratio may be 1.0, suggesting perfect overall calibration for the entire body of evidence⁶, even though there is much heterogeneity across studies. This can happen both for pooled *average calibration ratios*, and for pooled O:E ratios for specific risk strata or relevant subgroups.

Despite its analytical limitations in comparison to the calibration plot, the O:E ratio is more frequently available, and can be effectively pooled^{7,9}, which remains a problem for calibration plots and calibration curves. Also, the O:E ratios of the individual studies included in a systematic review can be displayed (and inspected) on a forest plot. Considering that there is yet no established methods to pool calibration plots, and that most systematic reviews in the field reports O:E ratios, the remainder of our discussion will focus on the use of the O:E ratio as a measure of calibration to be evaluated by the GRADE approach. The O:E ratio of interest might be that for the whole population combined, or in particular risk groups or covariate patterns.

Using GRADE to rate the certainty in model calibration

As with systematic reviews and meta-analyses of interventions, overall prognosis, prognostic factors, and diagnostic test accuracy, there is a need for guidance on determining certainty in inferences regarding predictive performance of prognostic models¹⁰⁻¹³. Although GRADE provides guidance for assessing the certainty in modelling the impact of interventions on health benefits and harms, and the economic efficiency of health interventions¹⁴, specific guidance for prognostic models remains necessary. We offer this GRADE concept paper to present insights conceived whilst developing guidance on appraising certainty of prognostic models.

Assessing certainty in a model's calibration creates unique challenges for the application of the GRADE approach. In this paper, we review such challenges and propose unique adaptations to the GRADE approach required for determining certainty in inferences regarding model calibration. We frame the discussion as four concepts that are necessary for rating certainty in the calibration of prognostic model studies. The concepts proposed in this paper are not necessary for rating model discrimination, and so discrimination is not discussed here. Measures of clinical utility (e.g., net benefit) and direct measures of impact (e.g., from randomised trials) are also beyond the scope of this paper.

Concept #1: The Target of Certainty Rating of Calibration in Systematic Reviews of Validation Studies of Prognostic Models

The Target of Certainty Rating in Prognostic Models

In reviews of validation studies of prognostic models, one must first clarify the level of calibration one considers satisfactory (for instance, one might consider an O:E ratio of 0.9 to 1.1 satisfactory - a standard that we will use in our subsequent examples), and then rate one's certainty in whether that standard has or has not been met. In considering the O:E ratio, we consider model performance satisfactory when most validation studies report a ratio falling in the selected range. Unsatisfactory model performance exists when most studies report an O:E ratio much above or below the selected range. Important deviation from an O:E of 1 is context specific, as it depends, amongst other things, on the event rate in the population. In deciding what level of miscalibration is unsatisfactory, reviewers may wish to consider the implications of misclassification of risk. For example, consider the decision regarding undergoing cardiac transplantation, where benefit is expected only in patients with a mortality rate over the next year of more than 20%. Appreciable under- or overestimation in the calibration curve around the 20% risk will result in crucial suboptimal decisions.

The pattern of model performance across validation studies will of course often show intermediate performance between clearly satisfactory and clearly unsatisfactory. The occurrence of these intermediate results undermines our certainty in the judgment of satisfactory or unsatisfactory model performance.

Limitations of the O:E Ratio

In systematic reviews and meta-analyses of model calibration there exist two reasons that a single pooled O:E ratio (average calibration) may represent a poor summary metric to assess model prognostic performance. First, in a single validation study, an apparently satisfactory overall calibration ratio may exist despite serious miscalibration. Second, a similar issue can arise at the study level in meta-analyses of O:E ratio. These issues are discussed at length above under *Measures of model calibration*.

For instance, a systematic review and meta-analysis by Ebell et al. pooled O:E ratios calculated from each primary study to report on the calibration of the CRB-65 score, a model for estimating mortality risk among patients with community acquired pneumonia¹⁵. The authors observed a pooled estimate of 1.04 (95% CI, 0.91 - 1.19) consistent with excellent overall calibration. Most studies, however, reported O:E ratios representing considerable over- or underestimation of risk (figure 1A) – the apparently satisfactory pooled O:E ratio resulted from the similar number of over and underestimates.

Given these limitations, one should view an individual study O:E ratio close to 1 as necessary but, because it may represent overestimation in one risk group and underestimation in another, as insufficient for assessing model performance. Fortunately, despite its limitations, viewing a forest plot of the O:E ratio in individual studies, and their confidence intervals, can be extremely informative in evaluating the certainty of evidence from a set of validation studies that have addressed a particular model. We will now illustrate how this is the case.

Concept #2: The Role of Inconsistency of Results Across Studies in Deciding Whether to Rate Certainty in Satisfactory or Unsatisfactory Model Performance

For interventions, overall prognosis and prognostic factors, and diagnosis, assessment of heterogeneity does not bear on decisions regarding the target of inference, but rather on certainty in the inferences that emerge: that is, inconsistency in results decreases confidence in inferences regarding intervention effects, prognostic power, or diagnostic accuracy. In contrast, for prognostic models the GRADE inconsistency domain bears a role in first determining the target of certainty assessments: the distribution of point estimates reported by individual validation studies, beyond differences across studies that may be observed by chance alone, can inform whether the model's performance is satisfactory or unsatisfactory. Large inconsistency (Table 1, example 4), or consistent results with an O:E ratio far from 1 (Table 1, example 6), will dictate rating one's certainty in an unsatisfactory model (and establish high certainty); consistent results near 1 (Table 1, example 1) will dictate

rating certainty in a satisfactory model (and will establish high certainty). Other results, as we will illustrate, will undermine certainty in inferences regarding satisfactory or unsatisfactory model performance.

To judge the extent of inconsistency, and thus whether to make the inference that model performance is satisfactory or unsatisfactory, one may examine the overlap of studies' point estimates and 95% CI, and statistical measures of heterogeneity. One may also assess inconsistency by generating 95% prediction interval for the O:E ratios⁶. Instances in which the prediction intervals are narrow and point estimates are near 1.0 provide reassurance that the model performance is satisfactory. Similarly, one may use tau-squared (the estimate of between-study variance) to directly inform presence of heterogeneity. In instance where tau-squared is 0, the differences across studies may be due to chance alone. If, however, tau-squared is > 0 , it may be suggestive of heterogeneity. Statistical measures of heterogeneity should be used in conjunction with visually inspection of forest plots.

Figure 1A presents estimates of the O:E ratio from the available validation studies summarized in the Ebell et.al. systematic review¹⁵. The forest plot demonstrates that the prognostic model substantially underestimates risk in some validation studies and overestimates risk in others. Applying our approach beginning with deciding whether one is rating certainty in a satisfactory or unsatisfactory model, this degree of heterogeneity leaves no doubt that we are dealing with an unsatisfactory model, as when applied in some populations the model will overestimate risk and in others underestimate and results will thus be untrustworthy. Please note that in making this inference we trust that authors pursue efforts to explore the possible sources of the observed heterogeneity – a requisite step in any optimal systematic review – failed. We will challenge this assumption later on (Concept #4).

Table 1 provides six hypothetical examples that further illustrate our approach. In the first three examples, the O:E ratio is sufficiently consistent with point estimates sufficiently near 1.0 that rating certainty in satisfactory performance is preferable. In the latter three, that is not the case: in examples 4 and 5 the degree of inconsistency in results is sufficient that one should rate certainty that the model performance in validation studies has proved it unsatisfactory; in the 6th, results are consistently far from an O:E ratio of 1.0.

Concept #3: The Role of Inconsistency of Results Across Studies in Judging our Certainty in the body of evidence

Once reviewers have considered heterogeneity in deciding whether to rate certainty in satisfactory (Table 1, examples 1 to 3) or unsatisfactory (Table 1, examples 4 to 6) model performance, they will then reconsider heterogeneity (along with risk of bias, precision, directness, and publication bias, to be discussed in future GRADE guidance on evaluating certainty in a body of evidence on validation of prediction models) in deciding on certainty of evidence following the standard GRADE approach.

Rating satisfactory performance of calibration. In the first example in Table 1 the studies are extremely consistent (not serious inconsistency) and all have O:E values near 1.0, posing no challenge to the inference that the model's performance is satisfactory in terms of overall calibration. In the second example, heterogeneity is sufficient to raise doubts regarding the inference of satisfactory performance; in our judgment, however, whether or not to rate down is a close call. Some reviewers may rate down for serious inconsistency, whereas others may not. In the third example, we are still rating our certainty in the model's satisfactory performance, but heterogeneity is sufficient to mandate rating down our overall certainty in the evidence for serious inconsistency. These first three examples show that we don't expect prognostic models to work well in all validation settings. Rather the judgment is that miscalibration in a few settings is deemed acceptable.

Rating unsatisfactory performance of calibration. In the fourth example almost all estimates of the O:E ratio are far from that corresponding to the initial model. The degree and nature of the inconsistency mandates rating certainty that the calibration is unsatisfactory and, with respect to inconsistency, leaves no doubt about this inference. This example highlights the key difference when one decides to rate certainty in the unsatisfactory performance of the model. Here – in contrast to GRADE for questions of intervention, overall and prognostic factors, and diagnostic test accuracy in which serious or very serious inconsistency always mandates rating down one's certainty – the large degree of inconsistency, if anything, bolsters certainty in the unsatisfactory performance of the model.

In the fifth example, results are less clear. There are sufficient studies in which, using the original model, the O:E ratio approximates 1.0, that our inference that the model is unsatisfactory becomes less secure. Due to this serious inconsistency (though most results suggest an unsatisfactory model – point estimates of the O:E ratio far from one - some point estimates are near the O:E value of 1.0), we rate down our certainty that the model's overall calibration is unsatisfactory. The actual results of the Ebell et. al. review in Figure 1A correspond most closely to this hypothetical situation.

Concept #4: The Distribution of the Point Estimates for O-E Ratios Bears on Implications for Further Research

In Attempting to explain the heterogeneity observed in the review by Ebell et al (Figure 1A), we conducted a *post hoc* subgroup analysis (figure 2). We classified individual studies based on overall observed risk of mortality: <5% (low observed risk), 5% to 10% (intermediate observed risk), and >10% (high observed risk). CRB-65 overestimated risk of mortality in studies with a low observed risk of mortality (O:E ratio, 0.54 [95% CI, 0.36 – 0.71]), was much better calibrated in studies with intermediate observed risk of mortality (O:E ratio, 0.93 [95% CI, 0.75 – 1.11]), and underestimated risk in studies with high observed risk of mortality (O:E ratio, 1.45 [95% CI, 1.34 – 1.55]).

Assuming that this subgroup analysis is credible (which remains to be established, given the post-hoc nature of our analysis and the convenient selection of thresholds¹⁶), and that clinicians could identify whether their patients belong a low, medium or high-risk population (an even more questionable assumption), in contrast with the interpretation of Figure 1A, these results would provide clear direction for future research.

In Figure 1A heterogeneity remains unexplained and results provide few if any clues to producing a more satisfactory model. Here, future studies must start from first principles, seeking new predictors that may better generalize across health care settings (in other words, back to the drawing board).

In Figure 1B, if one could identify that a patient comes from an intermediate risk group, one could apply the initial model with at least moderate certainty (assuming no problems with risk of bias, indirectness, or publication bias; problems with either inconsistency or precision do not appear severe). For low and high-risk groups, the initial model is unsatisfactory. The consistency of under- and overestimates in the risk groups suggests, however, that the predictors in the initial model may work well if recalibrated with a new appropriate baseline risk⁶. One might therefore use data from one of the validation studies to recalibrate the model and the others to validate the recalibrated model, anticipating satisfactory performance of the recalibrated model itself. This example shows the potential application of credible subgroup analysis using characteristics that clinicians can easily identify in their patients. One could apply the same logic to the hypothetical 6th example in Table 1.

Concluding remarks

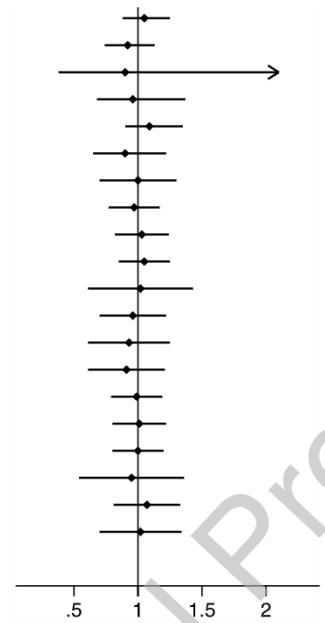
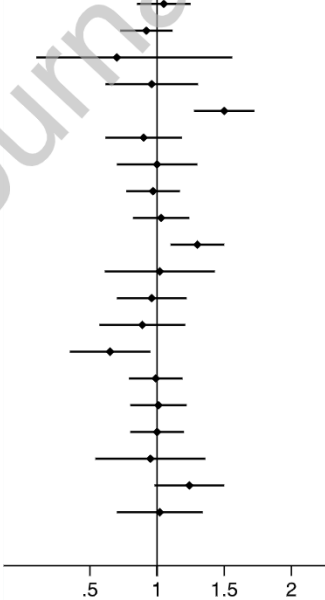
Calibration is a crucial measure to evaluate for a prognostic model¹⁷. Currently, review authors seldom conduct such careful evaluations, instead reporting on calibration performance using a variety of unsatisfactory approaches - or not at all¹⁸. Moreover, even review authors intent on producing optimal summaries may be limited by suboptimal or variable reporting of calibration in primary studies. Indeed, our GRADE project group, working on developing guidance for certainty ratings for prognostic models, has noted a paucity of systematic reviews that can serve as exemplars for developing our guidance; inconsistent and limited reporting in those that are available; the only recent development of a satisfactory risk of bias instrument for validation studies of models^{19 20}; and limitations related to risk specific calibration highlighted in this article.

The project group recognizes that clinical utility is best examined using net benefit measures, and results from impact studies⁸. However, the most common reported measure in primary studies hence systematic reviews thereof is the O:E statistic for the entire validation study population and, sometimes, risk strata based on predicted risks. Hence, in this paper we focused on adopting GRADE for certainty in calibration as assessed by O:E. This relates to the very weakest level of calibration assessment, but is still important to assess overall calibration to gain initial insight into the model's potential usefulness in clinical practice and shared decision making. Ideally more nuanced investigations would involve synthesis of calibration plots, and curves, but this likely requires the use of individual participant data meta-analyses.

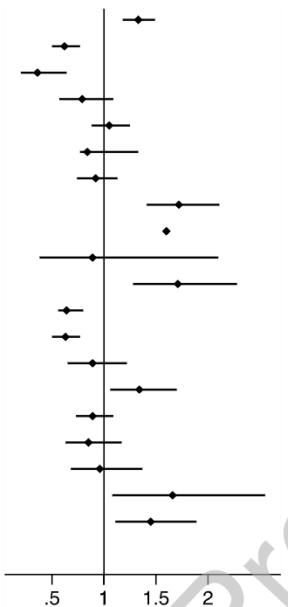
Methodological guidance by Debray et al. may help authors of systematic reviews improve data extraction and reporting practices for calibration⁷; adherence to the TRIPOD statement will improve reporting of calibration in primary studies²¹. Pending improvements in both individual validation studies and reviews summarizing these studies, and the consistent availability of impact studies, authors endeavoring to provide guidance regarding certainty of evidence from prognostic models may consider implementing suggestions we have provided in this paper.

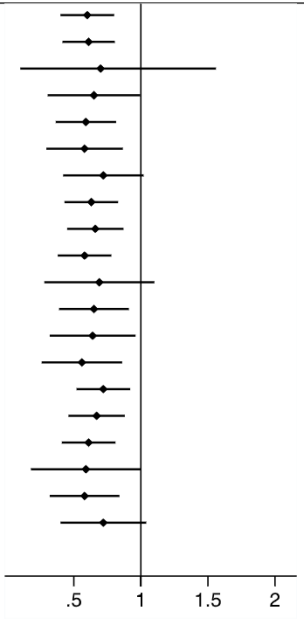
Nevertheless, in undertaking this work, our group has developed insights that distinguish assessing certainty in calibration assessment from validation studies of prognostic models from all previous GRADE guidance. These insights will not only prove useful in the project group completing its work, but in informing subsequent conceptual exploration of properties of prognostic models and inferences from evidence regarding their usefulness.

Table 1 – Degree of Inconsistency and Impact on Inferences Regarding Model Performance as measured by the O:E statistic in the entire population*

Possible inferences on the basis of degree of heterogeneity	Forest plots of O:E Ratios	Inference on model performance	Judgment about inconsistency
<p>1. Rating our certainty in satisfactory model performance: High certainty</p>		<p>Bearing in mind that the O:E measure is an insufficient measure on its own, we make the inference that the model is well calibrated as all studies agree that the O:E ratio is 1.0.</p>	<p>Not Serious</p> <p>We will not rate down for inconsistency as the model consistently works across all studies.</p>
<p>2. Rating certainty in satisfactory model performance: Possible rating down for inconsistency.</p>		<p>Amongst all included studies, most suggest that the O:E ratio is near 1.0. One can still make the inference that the model is well calibrated on average.</p>	<p>Not serious (some may judge as serious)</p> <p>In judging inconsistency, one may become more concerned. At this point, some authors or reviewers may conclude sufficient heterogeneity in average calibration for the patients studied to rate down for inconsistency. Some may not be</p>

			concerned and will not rate down.
3. Rating certainty in satisfactory model performance: Certainty rated down for inconsistency		In half of the studies, the reported O:E ratio suggests that the model is performing adequately. In the other half, it is not working well. One may still make the inference that the model is well calibrated on average.	<p>Serious, but explore in subgroup OR sensitivity analysis. If not explained: rate down by one level</p> <p>Due to the extensive observed heterogeneity, our certainty in the inference that the model works is decreased.</p> <p>In some settings / populations/ subgroups, the model performance is satisfactory but in other it is not.</p>
4. Rating certainty in unsatisfactory model performance: High certainty		In the following example, there is only one study with an O:E ratio of 1.0. In this instance, our inference would be that the model is poorly calibrated in the patients studied.	<p>Not serious</p> <p>This model is consistently miscalibrated (except for in one study). Therefore, we would not rate down for inconsistency. However, one cannot quantify the magnitude of miscalibration.</p>

<p>5. Rating certainty in unsatisfactory model performance: Rate down for inconsistency.</p>		<p>Some of the studies are scattered around an O:E ratio of 1.0. Enough studies are far from an O:E of 1.0, therefore we make the inference that the model is poorly calibrated in the patients studied.</p>	<p>Serious, but explore in subgroup OR sensitivity analysis. If not explained: rate down by one level</p> <p>In this example, although there were enough studies away from an O:E of 1.0 for us to make the inference that the model is miscalibrated, there are also enough studies close to an O:E of 1.0 for us to have doubt about our inference. Therefore, in this example, we would rate down for inconsistency.</p> <p>In some settings/ population/ subgroups the model performance is satisfactory whereas in others it is not.</p>
--	--	--	--

<p>6. Rating certainty in unsatisfactory model performance: High Certainty</p>		<p>We make the inference that the model is not well calibrated as all studies agree that the O:E ratio is less than 1.0.</p>	<p>Not Serious</p> <p>We will not rate down for inconsistency as the model consistently works across all studies.</p>
--	---	--	--

* In examples 2 to 6 we assume that tau-squared > 0, and thus that chance cannot explain variability.

Competing interest statement

All authors declare they did not receive support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years exist, nor do other relationships or activities that could appear to have influenced the submitted work. All authors are members of the GRADE working group.

Contributions

All authors contributed to the generation of the research hypothesis, participated to the discussion of its content and approved the final version of the manuscript. FF, and AI selected the systematic reviews used as examples and prepared summary of finding tables used in the process. FF drafted the manuscript, AI is the guarantor.

Ethical approval

Not applicable.

Funding

No external funding

Data sharing

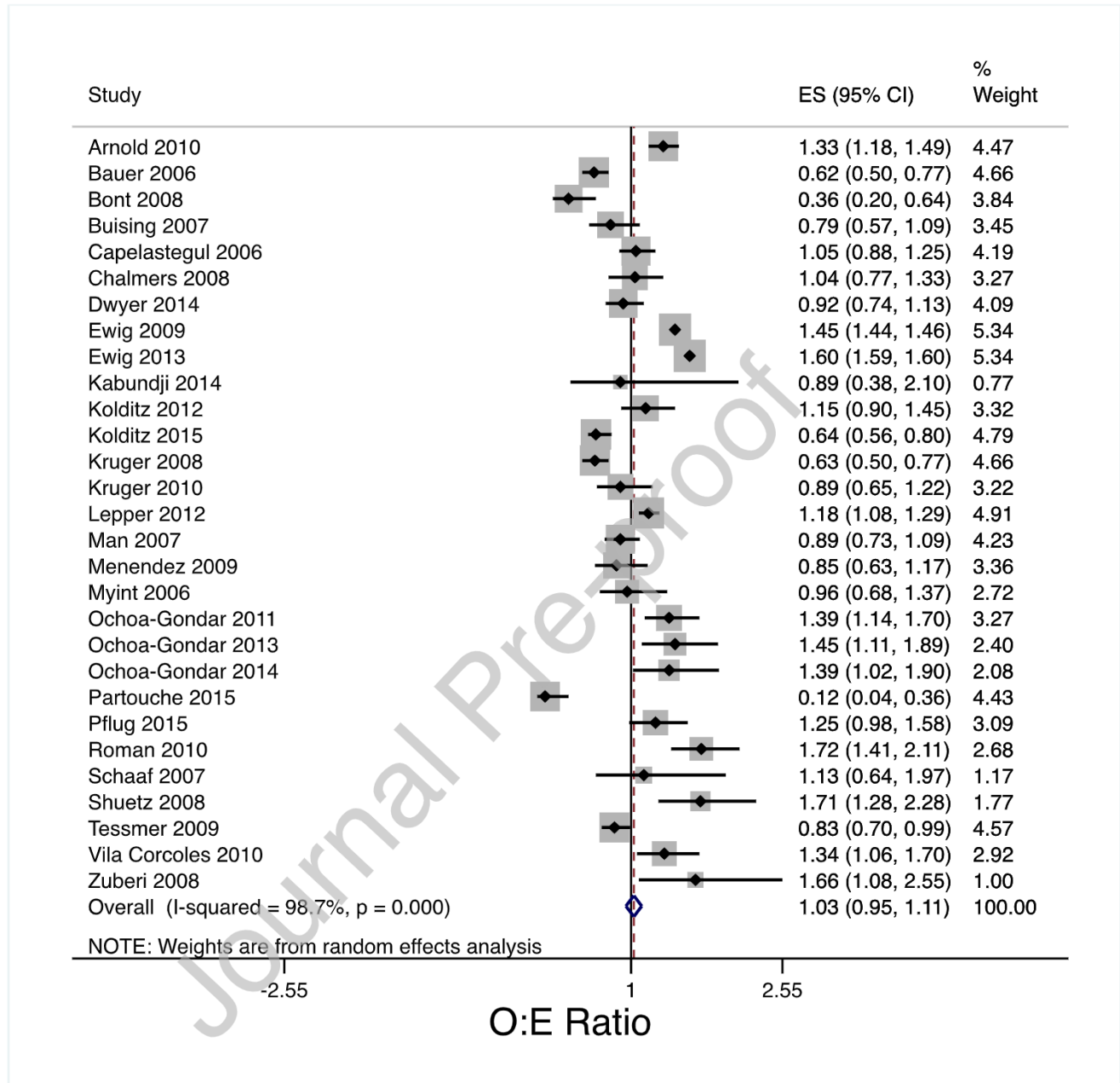
No additional data available.

References

1. Riley RD, van der Windt D, Croft P, et al. Prognosis Research in Health Care Concepts, Methods, and Impact: Concepts, Methods, and Impact: Oxford University Press 2019.
2. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10(2):e1001381. doi: 10.1371/journal.pmed.1001381 [published Online First: 2013/02/09]
3. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21(1):128-38.
4. 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 [published Online First: 2016/01/27]
5. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. *JAMA* 2017;318(14):1377-84. doi: 10.1001/jama.2017.12126 [published Online First: 2017/10/20]
6. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;353:i3140. doi: 10.1136/bmj.i3140 [published Online First: 2016/06/24]
7. 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 [published Online First: 2017/01/07]
8. Van Calster B, Nieboer D, Vergouwe Y, et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016; 74:167-76. doi: 10.1016/j.jclinepi.2015.12.005
9. Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res* 2019;28(9):2768-86. doi: 10.1177/0962280218785504 [published Online First: 2018/07/24]
10. Foroutan F, Guyatt G, Zuk V, et al. GRADE Guidelines 28: Use of GRADE for the assessment of evidence about prognostic factors: rating certainty in identification of groups of patients with different absolute risks. *J Clin Epidemiol* 2020;121:62-70. doi: 10.1016/j.jclinepi.2019.12.023 [published Online First: 2020/01/27]
11. Guyatt GH, Oxman AD, Schunemann HJ, et al. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol* 2011;64(4):380-2. doi: 10.1016/j.jclinepi.2010.09.011 [published Online First: 2010/12/28]
12. 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 [published Online First: 2015/03/18]

13. Schunemann HJ, Mustafa RA, Brozek J, et al. GRADE guidelines: 22. The GRADE approach for tests and strategies-from test accuracy to patient-important outcomes and recommendations. *J Clin Epidemiol* 2019;111:69-82. doi: 10.1016/j.jclinepi.2019.02.003 [published Online First: 2019/02/11]
14. Brozek JL, Canelo-Aybar C, Akl EA, et al. GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence-An overview in the context of health decision-making. *J Clin Epidemiol* 2021;129:138-50. doi: 10.1016/j.jclinepi.2020.09.018 [published Online First: 2020/09/28]
15. Ebell MH, Walsh ME, Fahey T, et al. Meta-analysis of Calibration, Discrimination, and Stratum-Specific Likelihood Ratios for the CRB-65 Score. *J Gen Intern Med* 2019;34(7):1304-13. doi: 10.1007/s11606-019-04869-z [published Online First: 2019/04/18]
16. Schandelmaier S, Briel M, Varadhan R, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. *CMAJ* 2020;192(32):E901-E06. doi: 10.1503/cmaj.200077 [published Online First: 2020/08/12]
17. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17(1):230. doi: 10.1186/s12916-019-1466-7 [published Online First: 2019/12/18]
18. Wessler BS, Lai Yh L, Kramer W, et al. Clinical Prediction Models for Cardiovascular Disease: Tufts Predictive Analytics and Comparative Effectiveness Clinical Prediction Model Database. *Circ Cardiovasc Qual Outcomes* 2015;8(4):368-75. doi: 10.1161/CIRCOUTCOMES.115.001693 [published Online First: 2015/07/15]
19. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med* 2019;170(1):W1-W33. doi: 10.7326/M18-1377 [published Online First: 2019/01/01]
20. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med* 2019;170(1):51-58. doi: 10.7326/M18-1376 [published Online First: 2019/01/01]
21. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Ann Intern Med* 2015;162(10):735-6. doi: 10.7326/L15-5093-2 [published Online First: 2015/05/20]

Figure 1A - Meta-analysis of O:E ratios as reported by Ebell et al.



Competing interest statement

All authors declare they did not receive support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years exist, nor do other relationships or activities that could appear to have influenced the submitted work. All authors are members of the GRADE working group.

Contributions

All authors contributed to the generation of the research hypothesis, participated to the discussion of its content and approved the final version of the manuscript. FF, VZ, and AI selected the systematic reviews used as examples and prepared summary of finding tables used in the process. FF drafted the manuscript, AI is the guarantor.

Ethical approval

Not applicable.

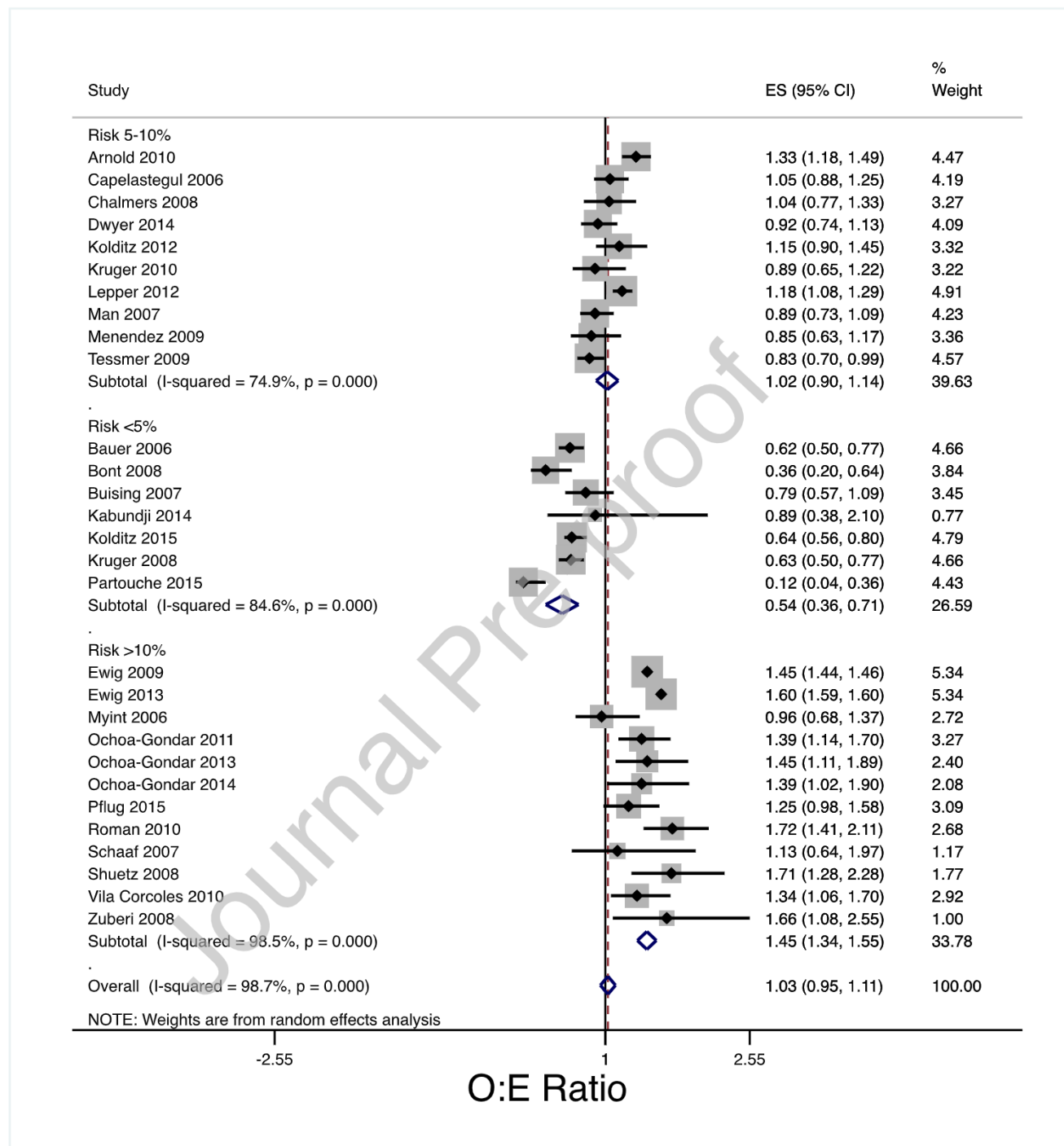
Funding

No external funding

Data sharing

No additional data available.

Figure 2 - Subgroup analysis of O:E based on overall cohort risk



Risk <5%: The observed risk of mortality in the overall cohorts was <5%; Risk 5 – 10%: The observed risk of mortality in the overall cohorts was between 5 to 10%; Risk >10%: The observed risk of mortality in the overall cohorts was >10%.

Author Statement

All authors contributed to the generation of the research hypothesis, participated to the discussion of its content and approved the final version of the manuscript. FF, and AI selected the systematic reviews used as examples and prepared summary of finding tables used in the process. FF drafted the manuscript, AI is the guarantor.

Journal Pre-proof