BMJ Open

# Empirical evidence on the impact of study characteristics and the performance of prediction models: a meta-epidemiological study



| | |
|---|---|
| Journal: | *BMJ Open* |
| Manuscript ID | bmjopen-2018-026160 |
| Article Type: | Research |
| Date Submitted by the Author: | 22-Aug-2018 |
| Keywords: | Meta-epidemiology, Prognosis, Prognostic models, Bias, Prediction |



Complete List of Authors:
Damen, Johanna; Cochrane Netherlands, University Medical Center Utrecht, Utrecht University; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University
Debray, Thomas; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University; Cochrane Netherlands, University Medical Center Utrecht, Utrecht University
Pajouheshnia, Romin; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University
Reitsma, Johannes; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University; Cochrane Netherlands, University Medical Center Utrecht, Utrecht University
Scholten, Rob; Cochrane Netherlands, University Medical Center Utrecht, Utrecht University; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University
Moons, Karel; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University; Cochrane Netherlands, University Medical Center Utrecht, Utrecht University
Hooft, Lotty; Cochrane Netherlands, University Medical Center Utrecht, Utrecht University; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University

# Empirical evidence on the impact of study characteristics and the performance of prediction models: a meta-epidemiological study

Johanna A A G Damen, Thomas P A Debray, Romin Pajouheshnia, Johannes B Reitsma, Rob J P M Scholten, Karel G M Moons, Lotty Hooft

Johanna A A G Damen
Assistant professor
Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Thomas P A Debray
Assistant professor
Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Romin Pajouheshnia
PhD fellow
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Johannes B Reitsma
Associate professor
Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Rob J P M Scholten
Professor

Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Karel G M Moons
Professor
Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Lotty Hooft
Associate professor
Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Correspondence to:
Johanna A A G Damen
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht
P.O. Box 85500
Str. 6.131
3508 GA Utrecht
The Netherlands
j.a.a.damen@umcutrecht.nl
+31 88 75 693 77

Word count: 4167

## Abstract

**Objectives:** To empirically assess the relation between study characteristics and prognostic model performance in external validation studies of multivariable prognostic models.

**Design:** Meta-epidemiological study.

**Data sources and study selection:** We searched electronic databases for systematic reviews of prognostic models published in the period 2010-2016. Reviews from non-overlapping clinical fields were selected if they reported common performance measures (either the concordance (c)-statistic or the ratio of observed over expected number of events (OE ratio)) from ten or more validations of the same prognostic model.

**Data extraction and analyses:** Study design features, population characteristics, methods of predictor and outcome assessment, and the aforementioned performance measures were extracted from the included external validation studies. Random effects meta-regression was used to quantify the association between the study characteristics and model performance.

**Results:** We included 10 systematic reviews, describing a total of 224 external validations, of which 221 reported c-statistics and 124 OE ratios. Associations between study characteristics and model performance were heterogeneous across systematic reviews. C-statistics were most associated with variation in population characteristics, outcome definitions and measurement, and predictor substitution. For example, validations with eligibility criteria comparable to the development study were associated with higher c-statistics compared to narrower criteria (difference in logit c-statistic 0.21 [95% CI 0.07, 0.35], similar to an increase from 0.70 to 0.74). Using a case-control design was associated with higher OE ratios, compared to using data from a cohort (difference in log OE ratio 0.97 [95% CI 0.38, 1.55], similar to an increase in OE ratio from 1.00 to 2.63).

**Conclusions:** Variation in performance of prognostic models across studies is mainly associated with variation in case-mix, study designs, outcome definitions and measurement methods, and predictor substitution. Researchers developing and validating prognostic models should realise the potential influence of these study characteristics on the predictive performance of prognostic models.

**Strengths and limitations of this study**

- To the best of our knowledge, this is the first meta-epidemiological study focusing on the association of study characteristics with estimates of prognostic model performance.
- We included all ten systematic reviews describing at least ten external validations of the same prognostic model, resulting in 224 external validations.
- We extracted relevant features of design and conduct according to existing checklists on quality assessment (CHARMS) and reporting of prediction model studies (TRIPOD).
- It was not feasible to fit multivariable meta-regression models due to the limited number of available, well-reported, validation studies within the individual reviews, rendering the effective sample size too small for multivariable meta-regression analyses.

## Introduction

Prediction models, including diagnostic and prognostic models, estimate the probability that an individual has or will develop a certain outcome (e.g. disease or complication). Hereto, they combine multiple predictors into an estimate of an individual's risk.[1] Before using a prediction model in clinical practice it is recommended to validate the performance of the model in a population other than the population in which the model was developed (so called external validation studies).[2] Such studies assess whether model predictions remain sufficiently accurate across different settings and populations. Obviously, it is important that the methodological quality of external validation studies is good, as otherwise estimates of the prediction model's performance may be biased and thereby lead to misleading conclusions on its generalizability to practice.

Systematic reviews have found that the performance of existing prediction models often varies substantially across external validation studies of those models.[3-5] These differences may not only appear due to random variation (when validation studies are small), but may also arise when model predictions are invalid because the model is applied in very different populations (eg, the association between predictors in the model and the outcome are different) or when design-related characteristics of the validation study (eg, measurement methods or variable definitions) are not well aligned with the original development study.[2 6]

To provide empirical evidence of the association of study characteristics with prediction model performance, a meta-epidemiological approach can be used. Studies using this approach have shown the influence of study characteristics on the effectiveness of interventions studied in randomized trials and on the accuracy of diagnostic tests.[7-12] For diagnostic prediction models evidence suggests estimates of performance may be biased in studies with certain study characteristics. One study found a higher diagnostic odds ratio in case-control studies, studies with differential outcome verification (ie, using different outcome assessments across study individuals), and with low sample size.[13] To date, no meta-epidemiological study has been performed investigating the possible impact of study characteristics on measures of the predictive performance of a prognostic model upon external validation, which is commonly quantified in terms of discrimination and calibration.[14] The aim of this study was to investigate sources of heterogeneity in the predictive performance of prognostic models. A meta-

epidemiological approach was used to synthesize evidence from a range of clinical fields. This study can serve as empirical evidence for design and analysis related bias in prognostic model studies.

## Methods

Search and selection of systematic reviews

We used an existing database (last updated on March 27, 2017) consisting of studies evaluating multiple existing prediction models, including narrative or systematic reviews of prediction models, or head-to-head comparisons of multiple prediction models validated on a specific dataset (See Supplement 1 for details of the search strategy). To construct this database, references identified by the search were screened for eligibility by one reviewer (GSC) on title, abstract and, if necessary, on full text. Subsequently, the full text of all articles in the database were screened for eligibility to the current project by another reviewer (JAAGD). We selected systematic reviews of prognostic models (ie, diagnostic models were excluded) that included at least ten studies that externally validated the same prognostic model, and that presented the performance of these models in terms of discrimination (concordance (c)-statistic or area under the receiver operating characteristic (AUC) curve), or calibration (observed expected (OE) ratio). Discrimination is the ability of the model to distinguish between people who will and who will not develop the outcome of interest, while calibration reflects the overall agreement between the total number of observed and predicted ('expected') events.[14] We excluded systematic reviews that selected studies based on specific study characteristics (eg, we excluded systematic reviews that did not include primary studies with a sample size below 100, if we were not able to identify the primary studies that had been excluded for this reason). Furthermore, we excluded reviews of prognostic models in which the weights of predictors in the original model were based on expert opinion rather than on coefficients estimated from a formal statistical approach. If more than one systematic review on the same prognostic model was identified, we included the one with the broadest inclusion criteria (eg, reviews focussing on specific patient populations were not preferred if a review with a broader population was available) or the most recent review (in this order of preference). When multiple prognostic models for the same condition were described in one systematic review which all fulfilled the selection criteria, we included the model with the highest number of external validations.

Selection of the primary external validation studies from the included systematic reviews

From the included systematic reviews we collected the primary studies in which the prognostic models were developed and externally validated. For primary external validation studies for which no measure

of discrimination (c-statistic) or calibration (OE ratio) was reported in the systematic review, we checked the full text of the primary study, and if performance was not reported, these studies were excluded. If primary external validation studies described multiple external validations of the same model and if there was no overlap in included participants between these external validations (eg, a model was validated in two different cohorts, or a model was validated in men and women separately), data were extracted for every external validation separately. If a model was validated multiple times on the same population (described in either one or multiple publications), we selected the external validation that was included in the systematic review. If the systematic review included all those external validations, we selected the one in which the study population and predicted outcome most closely resembled the population and outcome of the original model.

Data extraction and preparation

We extracted relevant features of design and conduct according to existing checklists on quality assessment (CHARMS) and reporting of prediction model studies (TRIPOD).[15-17] Information about study characteristics of studies in which the models were developed were extracted from the corresponding development papers. Information about study characteristics of primary external validation studies were first extracted from systematic reviews. This information was subsequently checked using the external validation studies and, if necessary, additional information was extracted by one reviewer (JAAGD or RP). Items we extracted included study type (eg, external validation only, development of a new model and external validation of a model), study design (eg, existing cohort, existing RCT), dependency of investigators (validation by independent investigators or investigators also involved in the development study), eligibility criteria for participant inclusion, setting, location (continent), study dates, number of centres, follow-up time and prediction horizon, age and gender distribution, deletion or substitution of predictors, outcome definition and measurement method, sample size and number of events, handling of missing data, and model performance (see Supplement 2 for details). The data extraction form was piloted on multiple articles by all reviewers (JAAGD, TPAD, LH, KGMM, RP, JBR, RJPMS).

For analysis purposes, some study characteristics had to be categorized or transformed (Supplement 2). For example, eligibility criteria of the validation study as compared to the development study had to be judged and categorized as comparable, narrower (if subgroups included in the development study were excluded from the validation study), broader (if subgroups excluded from the development study were included in the validation study), mixture (a combination of the two), or unclear. For setting, location,

predictors and outcome a similar categorization was used. If data on study characteristics were not reported in the primary external validation studies, these were either categorized as 'unclear' (in case of categorical study variables), or the study was excluded from the analyses of that (missing) study characteristic (in case of continuous study variables, such as sample size). In order to improve comparability between reviews, we standardized continuous study variables separately for each systematic review, i.e. for every variable we subtracted the mean and divided by the standard deviation of all external validations identified from the same systematic review.

Statistical analyses

We used a two-staged approach to study the possible association between study characteristics and predictive performance.

In the first stage, we fitted a univariable meta-regression model for every study characteristic within each systematic review with the logit c-statistic or log OE ratio as outcome variable.[18] The regression coefficients estimated from this meta-regression model indicate the difference in logit c-statistic or log OE ratio between a certain category of a study characteristic and a chosen reference category (ie, the category that was present in most systematic reviews) of that characteristic.

In the second stage, these regression coefficients were pooled by the use of a random effects model. This reflected the average influence of the study characteristic on model performance across all systematic reviews. For continuous characteristics, the regression coefficients obtained in the first stage were jointly pooled across reviews, using bivariate meta-analysis.[19 20] For categorical characteristics the results of univariable meta-analyses are presented. We planned to perform multivariable analyses to assess the association between various study characteristics in combination and the performance of prognostic models, but due to the paucity of data we were not able to do so. All analyses are described in more detail in Supplement 3.

Patient and public involvement

Patients and public were not involved in the design, recruitment or conduct of the study.

## Results

Identification and selection of studies

The search identified 2037 studies, of which 496 were included in the database and screened on full text, and 66 were further assessed (Figure 1). Finally, ten systematic reviews were included.[21-29] These reviews addressed external validations of the following prognostic models: ABCD2,[30] Essen Stroke Risk Score (ESRS),[31] EuroSCORE,[32] Framingham,[33] FRAX,[34] Injury Severity Score (ISS),[35] model for end-stage liver disease (MELD),[36] Pneumonia Severity Index (PSI),[37] Revised Cardiac Risk Index (RCRI),[38] and Simplified Acute Physiology Score (SAPS) 3[39] (Table 1). The reviews included 248 primary external validation studies with 274 external model validations (one study could describe multiple model validations). During data extraction, 73 of 274 validations were eventually excluded (most often for not reporting a performance measure), and 20 additional external model validations were identified (Figure 1). This resulted in the inclusion of 224 external validations, of which 221 could be included in the analyses of the c-statistic, and 124 in the analyses of the OE ratio. For the OE ratio, only validations of the EuroSCORE, Framingham, FRAX, PSI, RCRI and SAPS 3 prognostic models were included, due to the very low number of reported OE ratios in the validations studies for the other four prognostic models.

Description of included validations

The number of external validations per systematic review ranged from 11 to 30 (Table 1), and the median (IQR) sample size and number of events were 1069 (418-3043) and 92 (36-248), respectively. Most studies used an existing registry (N=104, 46%) or existing cohort (N=74, 33%) to validate the prognostic model. The median (IQR) c-statistic and OE ratio were 0.73 (0.64-0.82) and 0.92 (0.64-1.26), respectively. Predictive performance of the models was highly heterogeneous, even for external validations of the same prognostic model, as indicated by the wide prediction intervals (Table 1).
Not all information on the study characteristics was reported for all external validations (Table S1). Information was often unclear (eg, for outcome definitions (N=83, 37%) and handling of missing data (N=105, 47%)) or missing (eg, case-mix information such as mean age (N=28, 13%) and gender distribution (N=16, 7%)).

Discrimination

*Pooled models*

The pooled analyses across all systematic reviews (Figure 2, S1 and S2) showed that validation in a continent different from the development study was associated with a higher c-statistic, compared to validation in the same continent, and multicentre versus single centre validation studies were associated with a lower c-statistic. Comparable eligibility criteria for participant inclusion were also associated with higher c-statistics compared to narrower criteria, whereas a broader setting was associated with a lower c-statistic compared to a setting comparable to the development study. Although not statistically significant, validations with changes made to the predictors (ie, substitution or deletion of a predictor), or in which it was unclear whether all predictors were correctly measured, tended to have lower c-statistics compared to validations where no changes were made. In various reviews we found an association between the c-statistic and numerous other study characteristics, such as the study design, comparability of outcome definition, prediction horizon, sample size and number of events, and mean age of study participants (Figure 3, S2 and S3), only these were often not statistically significant when pooled together.

*Variation across reviews*

Across reviews we found associations of many study characteristics with the c-statistic although this was rather heterogeneous, and confidence intervals often overlapped (Figure 3 and Figure S3). For example, for study design, in six systematic reviews a higher c-statistic was found for validations that used an existing registry compared to an existing cohort, while in three reviews a lower c-statistic was found. In three systematic reviews we found a higher c-statistic in validations by independent investigators, while in five a lower c-statistic was found.

For other study characteristics, directions of associations were more consistent. For example, for most systematic reviews, validation studies with eligibility criteria narrower compared to the criteria used in the development study had a lower c-statistics while broader eligibility criteria were associated with higher c-statistics (Figure S3). C-statistics were also (slightly) higher in external validations with a setting comparable to the development study. Validation in a continent other than the development study in general was associated with a higher c-statistic, and multicentre studies had lower c-statistics compared to single centre studies. External validations in which it was unclear if there were changes made to the predictors had lower c-statistics (Figure S3).

Calibration

*Pooled analyses*

We found a significant association between study design and the OE ratio (Figure 4); using data from a case-control study (although known to be an inferior design for prognostic model research[16]) resulted in higher OE ratios, compared to using data from an existing cohort (though based on three external validations). Furthermore, higher OE ratios were found for studies in which the outcome was assessed by a panel of clinicians as compared to using a registry. In various reviews we found an association between the c-statistic and numerous other study characteristics, such as the duration of follow-up, year in which recruitment was started, sample size, standard deviation of age, and setting (Figure 4, S4, S5 and S6), only these were not statistically significant when pooled together.

*Variation across reviews*

For other categories of study design (other than the use of a case-control design), heterogeneous associations were found across systematic reviews (Figure 5). The associations of most other study characteristics with the OE ratio were also most often not consistent across systematic reviews (Figure S5 and S6). For example, for two systematic reviews external validations with appropriate handling of missing data had OE ratios closer to 1 compared to inappropriate handling of missing data, while in two reviews, OE ratios were further away from 1. Only for the continent in which the model was validated, directions were more consistent; OE ratios were closer to 1 if the continent was comparable to the development, compared to validations in different continents (Figure S6).

## Discussion

### Principal findings

Using a comprehensive meta-analytical approach, we studied the association between study characteristics of prognostic model validation studies and the estimated model performance across ten clinical domains. We focused on objective study characteristics that can be extracted from published reports. The reporting of the primary external validation studies was often incomplete and inadequate. Key study characteristics, such as outcome definitions, handling of missing data, and even model calibration estimates were infrequently reported. Still, we found associations between various study characteristics and a model's predictive performance. Changes in a model's predictive performance were notably found in relation to validation studies with a case-control (versus cohort) design, with differences in case-mix, in continent (in which the model is validated), in eligibility criteria, in clinical setting, in number of centres (included in the validation study), in differences in outcome definitions and assessments, and in predictor substitutions.

### Explanations, strengths and weaknesses

Based on findings in meta-epidemiological studies on the effect of study characteristics and the efficacy of interventions[7-10] and diagnostic test accuracy,[11 12] we anticipated to find more statistically significant associations between study characteristics and model performance across the included systematic reviews from different domains. Although we included every systematic review that described at least ten external validation studies of the same prognostic model, resulting in more than 200 validations from 10 reviews, our analyses appeared to still be hampered by relatively low numbers of external validations per systematic review, combined with poor reporting and substantial heterogeneity within and across systematic reviews.

Conceptually, there are many potential sources of heterogeneity in model performance, such as differences in population characteristics, predictor and outcome definitions and measurements, and in many aspects of the statistical analyses (eg, dealing with missing data, sample size and selective loss to follow up). All these characteristics may act in isolation but could also be related to each other. The individual strength of the association of one characteristic with model performance is ideally addressed by adopting multivariable (meta)-regression models with the observed model performance estimates of the validation studies as dependent variable and the characteristics of multiple design features as independent variables.[10 12] Unfortunately, this approach was not feasible here due to the limited

number of available, well-reported, validation studies within the individual reviews, rendering the effective sample size too small for multivariable meta-regression analyses.

A general limitation of all meta-epidemiological studies, is the possibility that the effect of a certain study characteristic differs across systematic reviews which may nullify the effect when pooled together.[40] We also found numerous conflicting associations between a study characteristic and the reported predictive performance measures across reviews that were cancelled out in the pooled analyses.

Also, it is possible that the effect caused by individual study characteristics is small and therefore difficult to detect. Moreover, there might be some misclassification of study characteristics, caused either by our misinterpretation of what is reported, or by a lack of reporting, which could have diluted the effects of the study characteristics. Indeed, the c-statistic is often considered to be an insensitive measure to quantify changes in model performance.[41-43] In previous simulation studies, the c-statistic and OE ratio appeared to be strongly influenced by case-mix differences,[14 44 45] which may mask the possible (smaller) effects from design-related characteristics. Other measures that are less sensitive to case-mix differences, such as the calibration slope, could, however, not be studied here simply because they were (almost) never reported in our retrieved studies, as was also shown previously.[3]

We found greater variation in the methods used by external validation studies *between* models than within validations of the *same* model. For example, multiple imputation is the preferred method for handling missing data in prediction modelling.[46 47] However, in the field of cardiovascular disease, it seems common to handle missing data by performing a complete case analysis, while in the field of mortality prediction in surgical patients, typically researchers fill in 'normal' values if a value is missing. Finally, given the explorative nature of our analyses to identify potential areas of further research, we did not correct for multiple testing, though we tried to minimize the number of exploratory analyses.

Comparison to previous research

Despite above considerations, our findings, ie, the trends in the associations between study characteristics and model performance measures (though not always statistically significant), are in agreement with various previous simulation studies in this field.[14 44 46-48] For example, we confirmed that studies with more variation in case-mix show higher c-statistics, and lower c-statistics when a predictor was omitted from the model. However, we found lower c-statistics in studies with a broader setting and when the number of centres in a study was higher.

We also found a higher OE ratio in studies with a case-control design. Both simulation studies and meta-epidemiological studies in the fields of diagnostic tests and (mainly diagnostic) prediction models, have shown biased effect measures in studies using a case-control design.[11-14] Further, we found that the OE ratio was influenced by the method of outcome assessment, in agreement with previous studies that showed that higher diagnostic odds ratios were found in studies with differential outcome verification.[13] We also expected to find lower OE ratios when the validation population differed from the development population (eg, in terms of case-mix).[14] We could not systematically confirm this across all reviews, likely caused by heterogeneity between systematic reviews as indicated by the wide confidence intervals. Finally, we could not fully confirm the association between sample size and model performance that was previously found,[13] although we found similar trends in part of the reviews.

Implications for future research

In agreement with many previously conducted systematic reviews on prediction models,[3 49-53] we still and again found poor reporting of prediction model studies. Meta-epidemiological studies of prediction model studies would highly benefit from complete reporting according to the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement.[16 17]

We also believe that more research is urgently needed to evaluate under which circumstances certain design choices may lead to heterogeneity in prediction model performance, and to incorporate these issues in the appraisal of prediction model studies. There is a need for more guidance on how to score items of critical appraisal checklists for prediction model studies, such as the CHARMS checklist.[15]

Several options exist to gain more empirical insight in design related bias in prediction model studies. Firstly, meta-epidemiological researchers can collect more external validation studies and try to correct for all issues that cause variation in performance of a model. We believe, however, that this is currently not feasible as we already included every systematic reviews describing at least ten validations of the same prognostic model. A second and much more efficient option is to collect the individual participant data (IPD) for all studies included in this review to directly study the effect of study characteristics on model performance.[54-58] Using IPD, it will also be possible to study different performance measures, like the case-mix adjusted c-statistic[44 59] and calibration slope.[14] Thirdly, new simulation studies could be performed to get more insight in design related bias in prediction model performance. Researchers could for example study the effect of using a different outcome definition or prediction horizon on the c-statistic of a model.

**Conclusion**

In this comprehensive meta-epidemiological study we found empirical evidence for an association between study characteristics and predictive performance of prognostic models. We found that predictive performance of prognostic models upon external validation is highly heterogeneous, but sensitive to various study characteristics, such as study design, case-mix, eligibility criteria, setting, methods of outcome definition and measurement, and predictor substitution. It is important that these characteristics are thus emphasized in the reporting and appraisal of prediction model studies. However, for a large part the observed heterogeneity in model performance remained unexplained, which is likely caused by the high number of factors that cause heterogeneity in predictive performance and may act in opposite directions whereas a multivariable meta-regression analysis across reviews simply was not possible.

Acknowledgments: The authors would like to acknowledge Prof. Gary S Collins (GSC) for building the database with systematic reviews of prediction models, which served as a basis for this paper.

Contributors: KGMM, JBR, TPAD, and LH conceived the study. All authors were involved in designing the study. JAAGD selected the articles. JAAGD and RP extracted the data. JAAGD analysed the data in close consultation with TPAD. All authors were involved in interpreting the data. JAAGD wrote the first draft of the manuscript which was revised by all authors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. JAAGD is guarantor.

Funding: Thomas Debray gratefully acknowledges the Netherlands Organization for Health Research and Development (grant number 91617050). Karel GM Moons received a grant from The Netherlands Organization for Scientific Research (ZONMW 918.10.615 and 91208004). The funder had no role in the design of the study; the collection, analysis, and interpretation of the data; or approval of the finished manuscript.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no 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; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Data sharing: no additional data are available.

Transparency: The lead author (JAAGD) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

## References


1. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012;98(9):683-90.
2. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2014.
3. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;353:i2416.
4. Alba AC, Agoritsas T, Jankowski M, Courvoisier D, Walter SD, Guyatt GH, et al. Risk prediction models for mortality in ambulatory patients with heart failure: a systematic review. *Circ Heart Fail* 2013;6(5):881-9.
5. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. *Breast Cancer Res Treat* 2012;132(2):365-77.
6. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ* 2009;338:b375.
7. Page MJ, Higgins JP, Clayton G, Sterne JA, Hrobjartsson A, Savovic J. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. *PLoS One* 2016;11(7):e0159267.
8. Berkman ND, Santaguida PL, Viswanathan M, Morton SC. AHRQ Methods for Effective Health Care. *The Empirical Evidence of Bias in Trials Measuring Treatment Differences*. Rockville (MD): Agency for Healthcare Research and Quality (US), 2014.
9. Savovic J, Jones H, Altman D, Harris R, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. *Health Technol Assess* 2012;16(35):1-82.
10. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 2008;336(7644):601-5.
11. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282(11):1061-6.
12. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174(4):469-76.
13. Ban JW, Emparanza JI, Urreta I, Burls A. Design Characteristics Influence Performance of Clinical Prediction Rules in Validation: A Meta-Epidemiological Study. *PLoS One* 2016;11(1):e0145779.
14. Steyerberg E. *Clinical prediction models: a practical approach to development, validation, and updating*: Springer Science & Business Media, 2008.
15. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11(10):e1001744.
16. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162(1):55-63.
17. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.

18. Snell KI, 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? *Stat Methods Med Res* 2017:962280217705678.
19. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2015.
20. 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.
21. Thompson DD, Murray GD, Dennis M, Sudlow CL, Whiteley WN. Formal and informal prediction of recurrent stroke and myocardial infarction after stroke: a systematic review and evaluation of clinical prediction models in a new cohort. *BMC Med* 2014;12:58.
22. Siregar S, Groenwold RH, de Heer F, Bots ML, van der Graaf Y, van Herwerden LA. Performance of the original EuroSCORE. *Eur J Cardiothorac Surg* 2012;41(4):746-54.
23. Marques A, Ferreira RJ, Santos E, Loza E, Carmona L, da Silva JA. The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis. *Ann Rheum Dis* 2015;74(11):1958-67.
24. Tohira H, Jacobs I, Mountain D, Gibson N, Yeo A. Systematic review of predictive performance of injury severity scoring tools. *Scand J Trauma Resusc Emerg Med* 2012;20:63.
25. Klein KB, Stafinski TD, Menon D. Predicting survival after liver transplantation based on pre-transplant MELD score: a systematic review of the literature. *PLoS One* 2013;8(12):e80661.
26. Chalmers JD, Mandal P, Singanayagam A, Akram AR, Choudhury G, Short PM, et al. Severity assessment tools to guide ICU admission in community-acquired pneumonia: systematic review and meta-analysis. *Intensive Care Med* 2011;37(9):1409-20.
27. Ford MK, Beattie WS, Wijeysundera DN. Systematic review: prediction of perioperative cardiac complications and mortality by the revised cardiac risk index. *Ann Intern Med* 2010;152(1):26-35.
28. Nassar AP, Malbouisson LM, Moreno R. Evaluation of Simplified Acute Physiology Score 3 performance: a systematic review of external validation studies. *Crit Care* 2014;18(3):R117.
29. Damen JAAG, Pajouheshnia R, Heus P, Moons KGM, Reitsma JB, Scholten RJPM, et al. Performance of the Framingham risk models and Pooled Cohort Equations for predicting 10-year risk of cardiovascular disease: a systematic review and meta-analysis. Manuscript submitted for publication.
30. Rothwell PM, Giles MF, Flossmann E, Lovelock CE, Redgrave JN, Warlow CP, et al. A simple score (ABCD) to identify individuals at high early risk of stroke after transient ischaemic attack. *Lancet* 2005;366(9479):29-36.
31. Diener HC, Ringleb PA, Savi P. Clopidogrel for the secondary prevention of stroke. *Expert Opin Pharmacother* 2005;6(5):755-64.
32. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg* 1999;16(1):9-13.
33. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97(18):1837-47.
34. Kanis JA, Oden A, Johnell O, Johansson H, De Laet C, Brown J, et al. The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. *Osteoporos Int* 2007;18(8):1033-46.
35. Baker SP, O'Neill B, Haddon W, Jr., Long WB. The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. *J Trauma* 1974;14(3):187-96.

36. Malinchoc M, Kamath PS, Gordon FD, Peine CJ, Rank J, ter Borg PC. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology* 2000;31(4):864-71.
37. Fine MJ, Auble TE, Yealy DM, Hanusa BH, Weissfeld LA, Singer DE, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med* 1997;336(4):243-50.
38. Lee TH, Marcantonio ER, Mangione CM, Thomas EJ, Polanczyk CA, Cook EF, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation* 1999;100(10):1043-9.
39. Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P, Campos RA, et al. SAPS 3--From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* 2005;31(10):1345-55.
40. Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med* 2002;21(11):1513-24.
41. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21(1):128-38.
42. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., 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(2):157-72; discussion 207-12.
43. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011;30(10):1105-17.
44. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010;172(8):971-80.
45. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. *BMJ* 2016;353:i3139.
46. Held U, Kessels A, Garcia Aymerich J, Basagana X, Ter Riet G, Moons KG, et al. Methods for Handling Missing Variables in Risk Prediction Models. *Am J Epidemiol* 2016;184(7):545-51.
47. Janssen KJ, Vergouwe Y, Donders AR, Harrell FE, Jr., Chen Q, Grobbee DE, et al. Dealing with missing predictor values when applying clinical prediction models. *Clin Chem* 2009;55(5):994-1001.
48. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol* 2012;12:82.
49. 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.
50. 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(3):268-77.
51. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9(5):1-12.
52. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak* 2006;6:38.
53. 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.
54. Debray TP, Koffijberg H, Vergouwe Y, Moons KG, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. *Stat Med* 2012;31(23):2697-712.

55. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* 2013;32(18):3158-80.
56. Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS Med* 2015;12(10):e1001886.
57. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, 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.
58. Riley RD, Price MJ, Jackson D, Wardle M, Gueyffier F, Wang J, et al. Multivariate meta-analysis using individual participant data. *Res Synth Methods* 2015;6(2):157-74.
59. White IR, Rapsomaniki E. Covariate-adjusted measures of discrimination for survival data. *Biom J* 2015;57(4):592-613.
60. Giles MF, Rothwell PM. Systematic review and pooled analysis of published and unpublished validations of the ABCD and ABCD2 transient ischemic attack risk scores. *Stroke* 2010;41(4):667-73.

Table 1: Description of included reviews and prediction models

| **Systematic review** | **Giles 2010**[60] | **Thompson 2014**[21] | **Siregar 2012**[22] | **Damen**[29] | **Marques 2015**[23] | **Tohira 2012**[24] | **Klein 2013**[25] | **Chalmers 2011**[26] | **Ford 2010**[27] | **Nassar 2014**[28] |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | ABCD2[30] | ESRS[31] | EuroSCORE[32] | Framingham[3] [3] | FRAX[34] | ISS[35] | MELD[36] | PSI[37] | RCRI[38] | SAPS 3[39] |
| Population | Patients with TIA | Adults with a previous CVD event | Adult patients who underwent cardiac surgery under cardiopulmonary bypass | Men without previous CHD event | General population | Injured patients | Patients with liver cirrhosis but without hepatocellular carcinoma who underwent elective transjugular intrahepatic portosystemic shunts | Inpatients with community-acquired pneumonia | Patients aged >=50 years who underwent nonemergent noncardiac procedures | ICU patients |
| Geographical location | United States and UK | Canada, United States, Europe | Europe | United States | Europe, Canada, Japan, United States, Australia | United States | United States | United States | United States | Worldwide |
| Patient recruitment | 1981-1998 | 1992-1995 | 1995 | 1971-1974 | 1980-1999 | 1968-1969 | 1991-1995 | 1989 | 1989-1994 | 2002 |
| Predicted outcome | Stroke | Recurrent ischemic stroke, MI and vascular death | Mortality | CHD | Osteoporotic fractures | All-cause mortality | All-cause mortality | 30-day hospital mortality | Major cardiac complications | Hospital mortality |
| Prediction horizon | 2 days | 1 year | 30 days | 10 years | 10 years | 3 months | 3 months | 30 days | 1 year | 90 days |
| **Performance development study** | | | | | | | | | | |
| C-statistic | 0.66 [95% CI 0.60, 0.71] | NR | 0.7875 | 0.74 | 0.63 | NR | NR | 0.84 | 0.759 [SE 0.032] | 0.848 |
| OE ratio | NR* | NR* | NR* | NR* | NR* | NR* | NR* | NR* | NR* | 1.00 [95% CI 0.98, 1.02] |
| ***Pooled* performance validation studies** | | | | | | | | | | |
| Number of external validations included in analyses | 16 | 11 | 22 | 23 | 30 | 34 | 14 | 24 | 23 | 27 |
| C-statistic [95% CI] | 0.66 [0.61, 0.71] | 0.60 [0.58, 0.62] | 0.79 [0.77, 0.81] | 0.68 [0.65, 0.71] | 0.66 [0.63, 0.68] | 0.86 [0.83, 0.88] | 0.64 [0.59, 0.68] | 0.80 [0.77, 0.82] | 0.69 [0.65, 0.72] | 0.83 [0.80, 0.85] |
| 95% PI | [0.54, 0.77] | [0.57, 0.63] | [0.74, 0.83] | [0.56, 0.78] | [0.54, 0.76] | [0.62, 0.96] | [0.48, 0.77] | [0.64, 0.89] | [0.53, 0.81] | [0.66, 0.92] |
| OE ratio [95% CI] | NA | NA | 0.54 [0.42, 0.68] | 0.58 [0.45, 0.76] | 1.10 [0.83, 1.47] | NA | NA | 0.94 [0.83, 1.06] | 2.70 [1.72, 4.25] | 0.89 [0.77, 1.03] |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 95% PI | NA | NA | [0.19, 1.51] | [0.20, 1.74] | [0.31, 3.93] | NA | NA | [0.55, 1.60] | [0.35, 20.75] | [0.42, 1.91] |

TIA: transient ischaemic attack, CVD: cardiovascular disease, CHD: coronary heart disease, ICU: intensive care unit, UK: United Kingdom, MI: myocardial infarction, NR: not reported, CI: confidence interval, PI: prediction interval, NA: not assessed.

*As the models are optimally fit in the development dataset, all OE ratios should be close to 1.

**Figure legends**

Figure 1: Flow chart of study selection.
SR: systematic review, IPD: individual participant data, MA: meta-analysis, NR: not reported, c: concordance, OE: observed expected.

Figure 2: Associations between study characteristics and logit c-statistic with regard to a reference category across 221 external validation studies and 10 different prediction models. Figure S1 shows these differences on the original scale if we assume a c-statistic of 0.70 in the reference category. For example, for comparability of eligibility criteria, if we assume a c-statistic of 0.70 in the reference category (narrower), this would result in c-statistics of 0.74 [0.72, 0.77], 0.73 [0.66, 0.79], 0.77 [0.68, 0.84], and 0.77 [0.59,0.89] in the categories comparable, mixture, broader, and unclear, respectively.

Figure 3: C-statistic for categories of study design, pooled using univariable meta-regression analyses within each systematic review. N represents the number of external validation studies in a specific category. C diff represents the difference in c-statistic with regard to a reference category (indicated with 'ref').

Figure 4: Associations between study characteristics and ln OE ratio with regard to a reference category across 124 external validation studies and 6 different prediction models. Figure S4 shows these differences on the original scale if we assume an OE ratio of 1.00 in the reference category. For example, for comparability of eligibility criteria, if we assume an OE ratio of 1.00 in the reference category (narrower), this would result in OE ratios of 0.83 [0.66, 1.05], 1.11 [0.72, 1.70], and 0.92 [0.54, 1.58] in the categories comparable, mixture, and broader, respectively.

Figure 5: OE ratio for categories of study design, pooled using univariable meta-regression analyses within each systematic review. N represents the number of external validation studies in a specific category. OE diff represents the difference in OE ratio with regard to a reference category (indicated with 'ref').

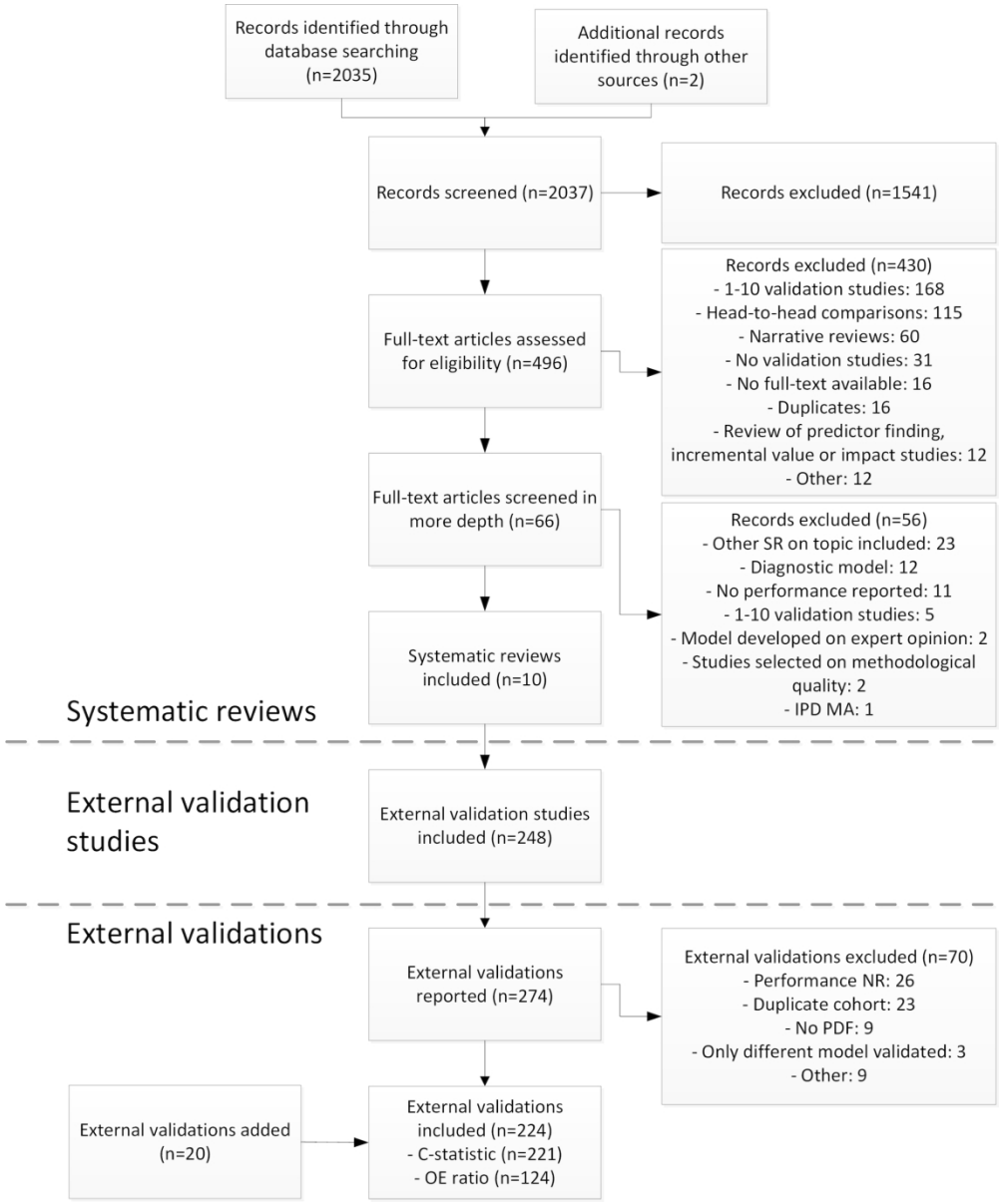



Figure 1: Flow chart of study selection.
SR: systematic review, IPD: individual participant data, MA: meta-analysis, NR: not reported, c: concordance, OE: observed expected.

184x222mm (300 x 300 DPI)

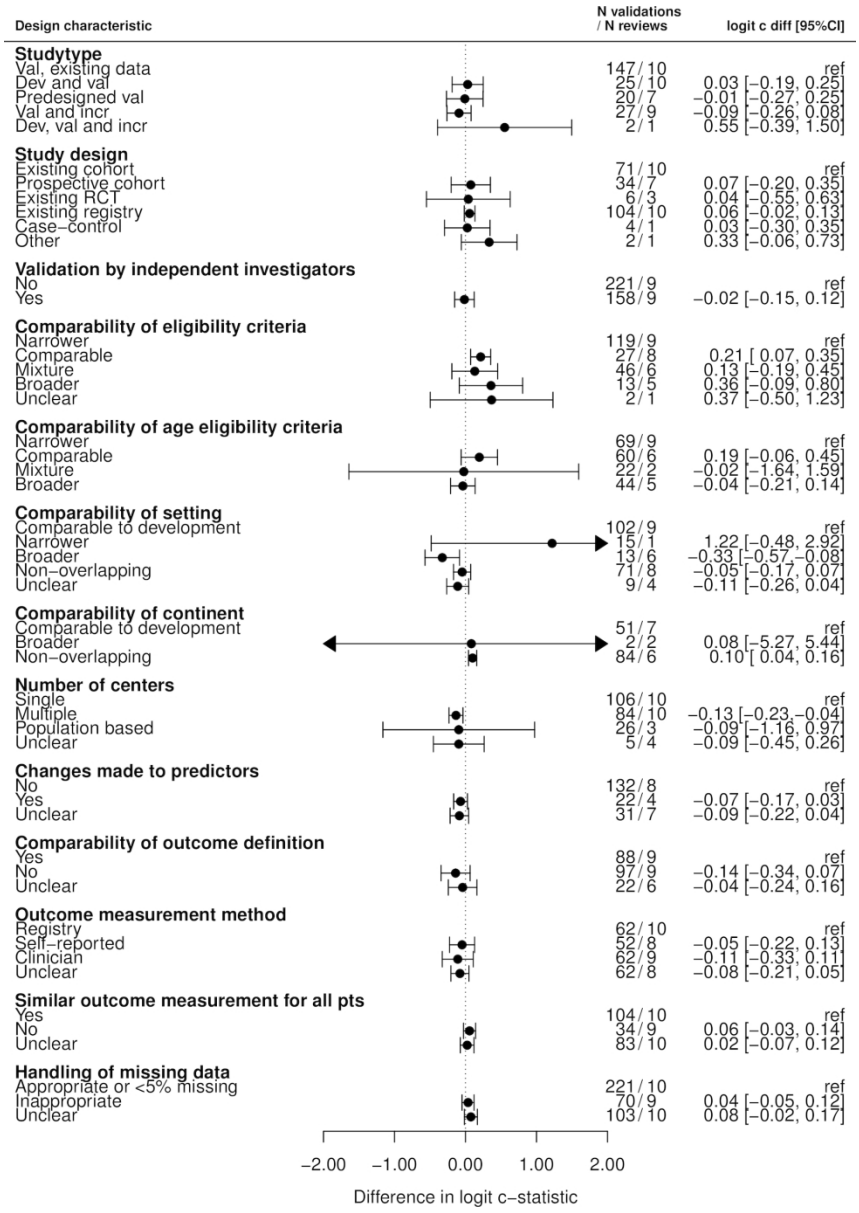



Figure 2: Associations between study characteristics and logit c-statistic with regard to a reference category across 221 external validation studies and 10 different prediction models. Figure S1 shows these differences on the original scale if we assume a c-statistic of 0.70 in the reference category. For example, for comparability of eligibility criteria, if we assume a c-statistic of 0.70 in the reference category (narrower), this would result in c-statistics of 0.74 [0.72, 0.77], 0.73 [0.66, 0.79], 0.77 [0.68, 0.84], and 0.77 [0.59,0.89] in the categories comparable, mixture, broader, and unclear, respectively.

157x222mm (300 x 300 DPI)

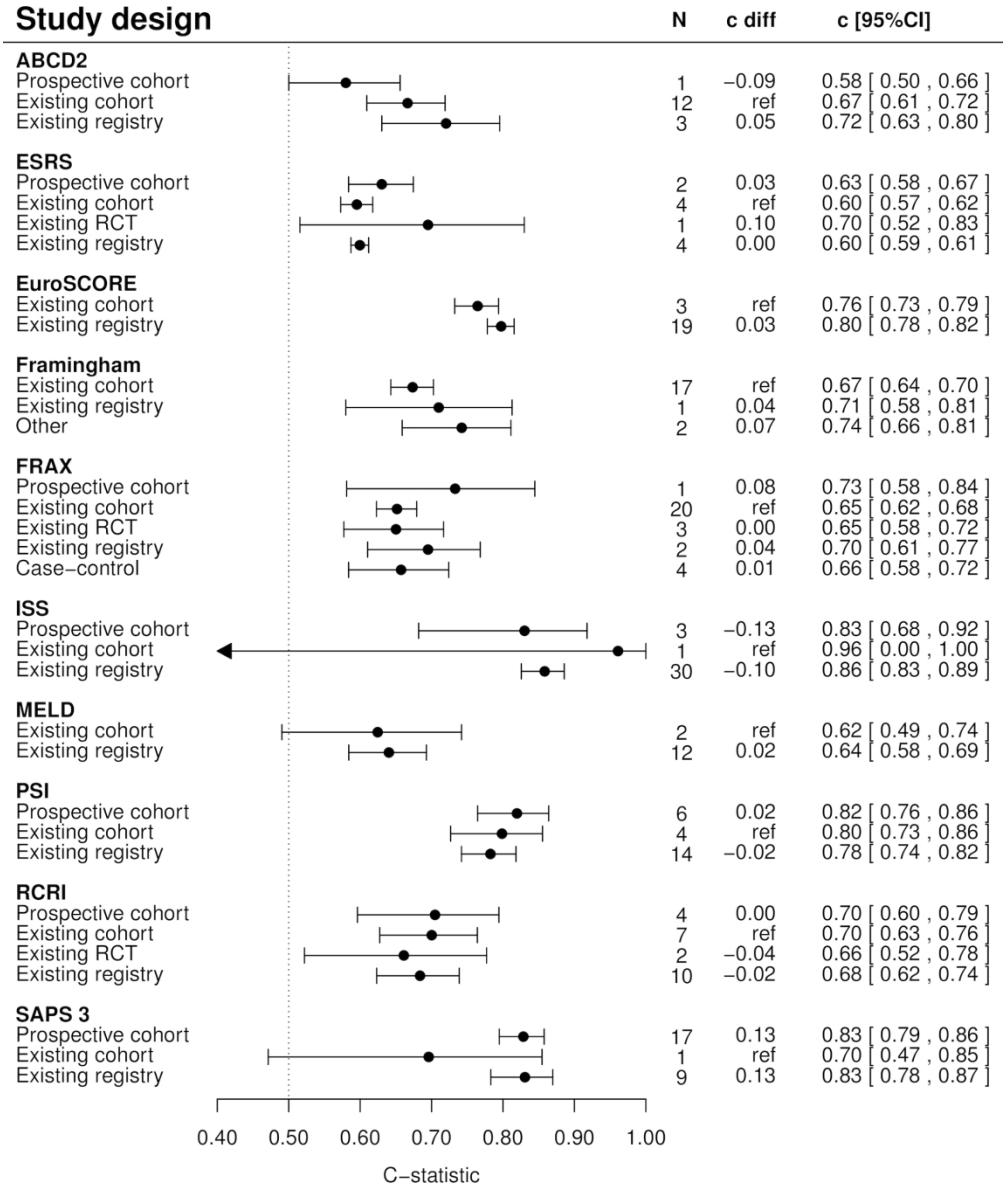



Figure 3: C-statistic for categories of study design, pooled using univariable meta-regression analyses within each systematic review. N represents the number of external validation studies in a specific category. C diff represents the difference in c-statistic with regard to a reference category (indicated with 'ref').

188x222mm (300 x 300 DPI)

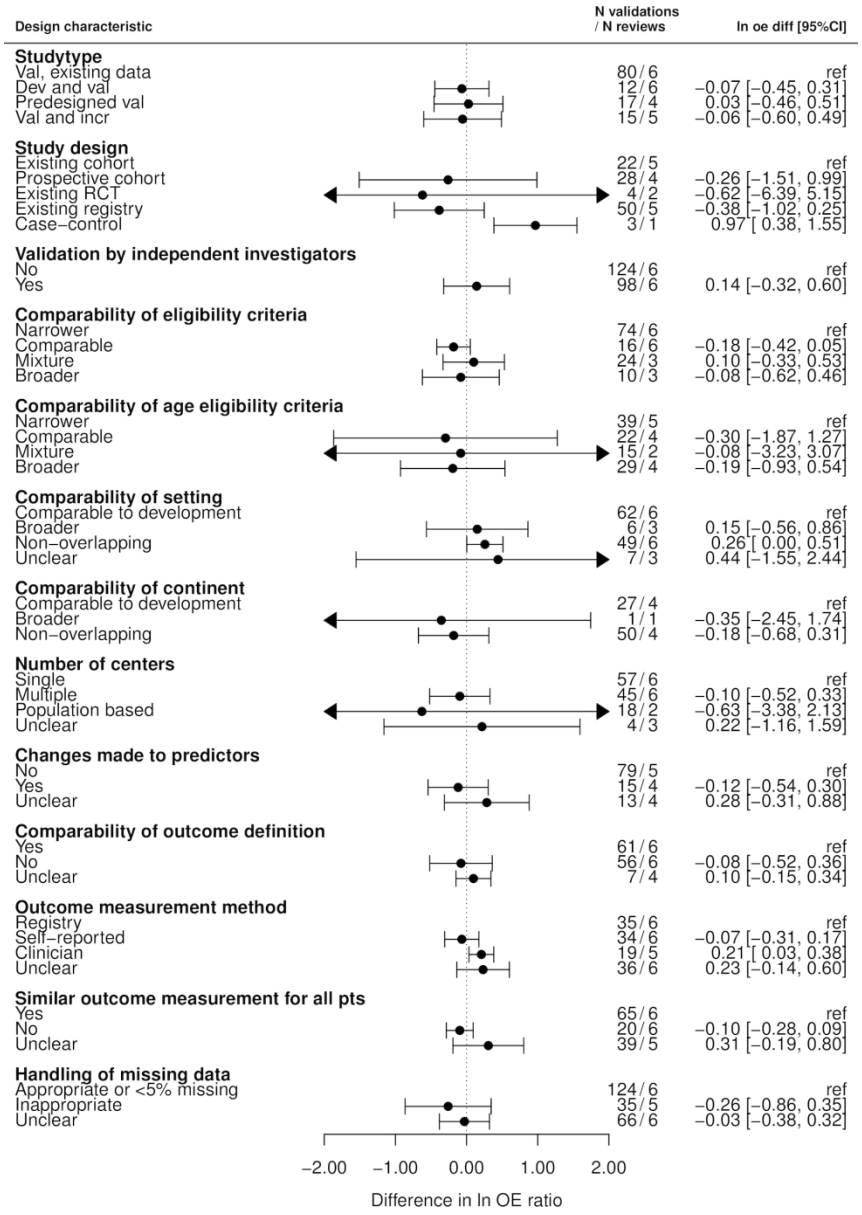



Figure 4: Associations between study characteristics and ln OE ratio with regard to a reference category across 124 external validation studies and 6 different prediction models. Figure S4 shows these differences on the original scale if we assume an OE ratio of 1.00 in the reference category. For example, for comparability of eligibility criteria, if we assume an OE ratio of 1.00 in the reference category (narrower), this would result in OE ratios of 0.83 [0.66, 1.05], 1.11 [0.72, 1.70], and 0.92 [0.54, 1.58] in the categories comparable, mixture, and broader, respectively.

158x222mm (300 x 300 DPI)

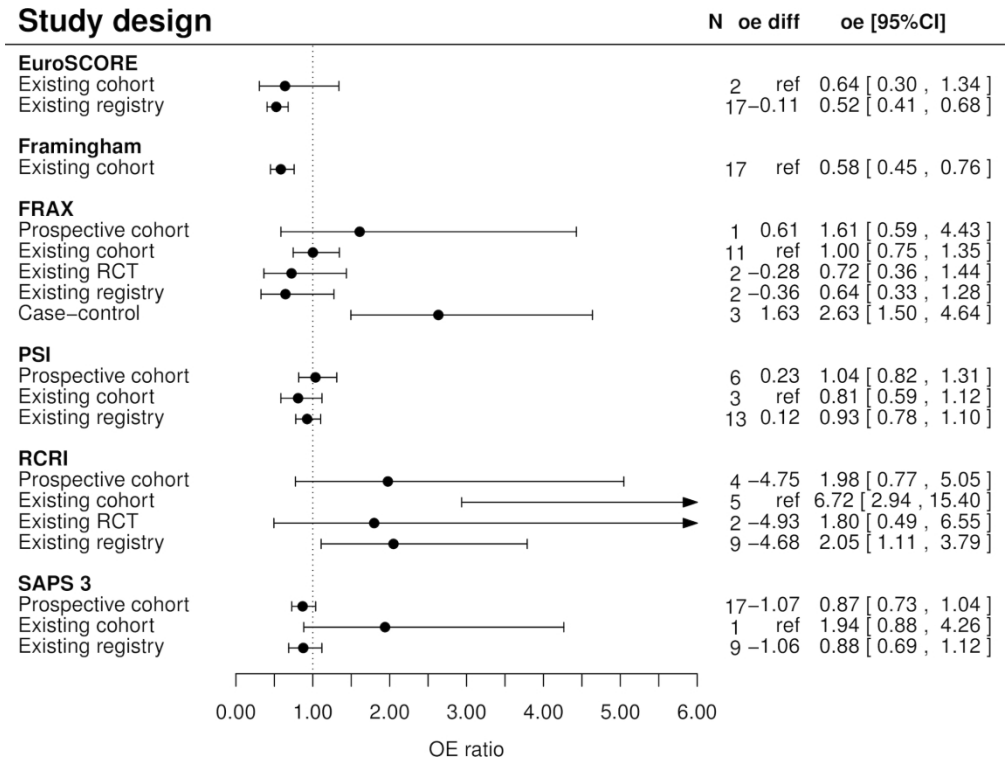



Figure 5: OE ratio for categories of study design, pooled using univariable meta-regression analyses within each systematic review. N represents the number of external validation studies in a specific category. OE diff represents the difference in OE ratio with regard to a reference category (indicated with 'ref').

190x142mm (300 x 300 DPI)

**Content**

**Supplement 1: Search string**

(clinical prediction[ti] OR
risk calculator*[ti] OR
risk index[ti] OR
risk indices[ti] OR
risk model*[ti] OR
risk prediction[ti] OR
risk score*[ti] OR
risk stratification[ti] OR
predictive model*[ti] OR
prediction model*[ti] OR
prediction rule*[tiab] OR
prognostic index[ti] OR
prognostic indices[ti] OR
prognostic model*[ti] OR
scoring system*[ti]) AND
(review[Publication Type] OR
review[ti] OR
critical appraisal[ti] OR
Bibliography[Publication Type] OR
Meta-analysis[Publication Type]) NOT
(Editorial[Publication Type] OR
Letter[Publication Type] OR
News[Publication Type])

**Supplement 2: Description of items extracted from studies and included in analyses**

| Item | Extracted from studies | Categorization / handling in analyses | Description / examples |
|---|---|---|---|
| Validated model | ABCD2, ESRS, EuroSCORE, Framingham Wilson, FRAX, ISS, MELD, PSI, RCRI, SAPS 3 | - | - |
| Study type | Predesigned validation study | Predesigned validation study | Study designed with the aim of validating a prediction model |
| | Validation study using existing data | Validation study using existing data | Study in which a prediction model is validated using a dataset collected for a different purpose than validating the model |
| | Development of new model and validation of different model | Development of new model and validation of different model | Study in which a model is developed and a model is validated |
| | Validation and incremental value | Validation and incremental value | Study in which a model is validated and in which the added value of one or more predictors is assessed |
| | Development, validation, and incremental value study | Development, validation, and incremental value study | Combination of the two above |
| Independent investigators | Yes | Yes | None of the authors of the development study was listed as author in the external validation study |
| | No | No | One or more of the authors of the development study was listed as author in the external validation study |
| Study design | Prospective cohort | Prospective cohort | |
| | Existing cohort | Existing cohort | |
| | Existing RCT | Existing RCT | |
| | Existing registry / medical records | Existing registry | |
| | Case-control | Case-control | |
| | Other (specify) | Other | |
| Eligibility criteria for participants | Copy/paste eligibility criteria of validation study | Comparable | Eligibility criteria comparable to development study |

| Item | Extracted from studies | Categorization / handling in analyses | Description / examples |
|---|---|---|---|
| | | Narrower | People included in the development study excluded in the validation study |
| | | Broader | People excluded in the development stud included in the validation study |
| | | Mixture | Combination of narrower and broader |
| | | Unclear | |
| Setting | Primary care<br>Secondary care<br>Tertiary care<br>Population based<br>Screening<br>Mixed<br>Unclear | Comparable | Same setting as development study |
| | | Broader | Same setting as development study, and participants from additional settings recruited |
| | | Non-overlapping | Setting in development study differs from validation study |
| | | Unclear | |
| Study dates | Start year of recruitment<br>End year of recruitment | Continuous, standardized per systematic review | |
| Prediction horizon | Time period for which predictions were made, eg, 10 years. | Continuous, standardized per systematic review | |
| Geographical location | Country and continent | Comparable | Model validated in the same continent as the development study |
| | | Broader | Model validated in the same and additional continents as the development study |
| | | Non-overlapping | Model validated in a different continent than the development study |
| Number of centres | Number of centres (numerical) | Single | |
| | | Multiple | |
| | | Population based | Participants not recruited at medical centres, but, for example, from a specific geographic area (eg, all individuals living in Framingham, US) |
| | | Unclear | |
| Case-mix: age mean | Mean and SD of age of | Continuous, standardized per | |

| Item | Extracted from studies | Categorization / handling in analyses | Description / examples |
|---|---|---|---|
| and sd | participants included in the study, or other available information about age distribution | systematic review | |
| Case-mix: gender | Percentage of men included in a study | Continuous, standardized per systematic review | |
| Predictors | Were predictors deleted from the model, or were predictors substituted with different predictors. | Yes | Changes made to predictors |
| | | No | No changes made to predictors |
| | | Unclear | |
| Predicted outcome | Full definition, including ICD-codes | Comparable | Outcome definition comparable to development study |
| | | Not comparable | Outcome definition not comparable to development study |
| | | Unclear | |
| Outcome - measurement method | Measurement method (eg, self-reported, interviews, expert panel), differences in outcome measurement between participants in the study | Yes | Outcome measurement similar for all participant |
| | | No | Systematic differences in outcome measurement between participants |
| | | Unclear | |
| Missing data | Number of participants with missing data, method of handling missing data | Appropriate | Missing data handled using multiple imputation, or <5% missing data (arbitrary cut-off) |
| | | Inappropriate | Missing data not handled using multiple imputation (eg, complete-case analysis, mean imputation), and >=5% missing data |
| | | Unclear | Unclear handling of missing data, and >=5% missing data |
| Number of participants | | Continuous, standardized per systematic review | |
| Number of events | | Continuous, standardized per systematic review | |

| Item | Extracted from studies | Categorization / handling in analyses | Description / examples |
|---|---|---|---|
| Model updating | Was the model altered before validating it, eg, using intercept recalibration. | NA | |
| Performance - c-statistic | C-statistic, AUC, 95% confidence intervals or SE | Logit transformation[1] | |
| Performance - OE ratio | OE ratio, predicted risks, presence of calibration plots or tables, 95% confidence intervals or SE | Ln transformation[1] | |

SD: standard deviation, NA: not applicable, C-statistic: concordance statistic, AUC: area under the receiver operating curve, SE: standard error, OE ratio: observed expected ratio.

Information regarding c-statistics and OE ratios when not reported was sometimes restored from other information reported in the paper. If the precision of the c-statistic was not reported, we estimated this from the c-statistic and sample size of the study, using the formula described by Newcombe and Hanley.[2 3] Various equations were used to estimate the standard error of the OE ratio, depending on which information was reported. All equations (as numbered) are described in the appendix of Debray et al.[4] If the SE of the OE ratio was reported, we used equation 16 to estimate the SE of ln(OE), if the observed event risk (Po), the expected event risk (Pe), and the SE of Po were reported, we used equation 51, and if only Po and Pe were reported we used equation 27.

**Supplement 3: Statistical analyses**

First we pooled the total OE ratio and c-statistic within each systematic review. Based on previous recommendations,[1 4] we pooled the log OE ratio and logit c-statistic using random-effects meta-analysis accounting for the presence of between-study heterogeneity, weighted by the inverse of the variance. We calculated 95% confidence intervals (CI) and (approximate) 95% prediction intervals (PI) to quantify uncertainty and the presence of between-study heterogeneity. The Hartung-Knapp-Sidik-Jonkman (HKSJ) method was used when calculating 95% CIs.[5] The 95% PI was calculated using the equation described previously.[4] The CI indicates the precision of the summary performance estimate and the PI provides boundaries on the likely performance in future model validation studies that are comparable to the studies included in the meta-analysis, and can thus be seen as an indication of model generalizability.[6]

To study the possible association between study characteristics and predictive performance, we used a two-stepped approach. In the first stage, we fitted a univariable meta-regression model (ie, a separate model for every study characteristic) within every systematic review, with the logit c-statistic or log OE ratio as outcome variable. This model was fitted with intercept term. Therefore, the effect estimates obtained from this meta-regression model indicate the difference in logit c-statistic or log OE ratio between a certain category of a study characteristic and a chosen reference category of that characteristic. As a reference category, we chose the category that was present in the highest number of systematic reviews allowing the inclusion of as many data as possible.

In the second stage, these effect estimates were pooled with a random effects meta-analysis model. This reflected the influence of the study characteristic on model performance over all systematic reviews. For continuous study characteristics, the intercept term and beta-coefficient from the first stage were jointly pooled across reviews using bivariate meta-analysis.[4 6] For categorical study characteristics the data available were not sufficient for the complexity of a multivariate model, so every category was pooled in a separate (univariate) meta-analysis.

As the estimates obtained with this approach are on the transformed scale (ie, the difference in logit c-statistic or log OE ratio between one category and the reference category), we transformed these back assuming a c-statistic of 0.7 or an OE ratio of 1.00 in the reference category. Also, we performed a second analysis where we again fitted a univariable meta-regression model, with the logit c-statistic or log OE ratio as outcome variable, but now without intercept term. This analysis enables the calculation

of an effect estimate for every category of a study characteristic and to back transform this to the original scale, yielding a pooled c-statistic or pooled OE ratio for each category of a study characteristic. We planned to perform multivariable analyses to assess the association between various study characteristics in combination and the performance of prediction models, but due to the paucity of data we were not able to do so.

All analyses were performed in R version 3.3.2,[7] using the packages metafor,[8] mvmeta,[9] metamisc,[10] and lme4.[11]

Table S1: Description of study characteristics and quality of reporting within each systematic review
Categorical variables

| | ABCD2 | ESRS | EuroSCORE | Framingham | FRAX | ISS | MELD | PSI | RCRI | SAPS 3 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Studytype** | | | | | | | | | | |
| Validation study using existing data | 9 (56%) | 7 (64%) | 21 (95%) | 18 (78%) | 26 (87%) | 24 (71%) | 10 (71%) | 16 (67%) | 11 (48%) | 8 (30%) |
| Development of new model and validation of different model | 2 (12%) | 2 (18%) | 1 (5%) | 2 (9%) | 1 (3%) | 4 (12%) | 3 (21%) | 2 (8%) | 2 (9%) | 6 (22%) |
| Development, validation, and incremental value study | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 2 (6%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Predesigned validation study | 1 (6%) | 1 (9%) | 0 (0%) | 0 (0%) | 1 (3%) | 1 (3%) | 0 (0%) | 5 (21%) | 1 (4%) | 10 (37%) |
| Validation and incremental value | 4 (25%) | 1 (9%) | 0 (0%) | 3 (13%) | 2 (7%) | 3 (9%) | 1 (7%) | 1 (4%) | 9 (39%) | 3 (11%) |
| **Study design** | | | | | | | | | | |
| Existing cohort | 12 (75%) | 4 (36%) | 3 (14%) | 20 (87%) | 20 (67%) | 1 (3%) | 2 (14%) | 4 (17%) | 7 (30%) | 1 (4%) |
| Prospective cohort | 1 (6%) | 2 (18%) | 0 (0%) | 0 (0%) | 1 (3%) | 3 (9%) | 0 (0%) | 6 (25%) | 4 (17%) | 17 (63%) |
| Existing RCT | 0 (0%) | 1 (9%) | 0 (0%) | 0 (0%) | 3 (10%) | 0 (0%) | 0 (0%) | 0 (0%) | 2 (9%) | 0 (0%) |
| Existing registry | 3 (19%) | 4 (36%) | 19 (86%) | 1 (4%) | 2 (7%) | 30 (88%) | 12 (86%) | 14 (58%) | 10 (43%) | 9 (33%) |
| Case-control | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 4 (13%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Other | 0 (0%) | 0 (0%) | 0 (0%) | 2 (9%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| **Validation by independent investigators** | | | | | | | | | | |
| No | 6 (38%) | 3 (27%) | 3 (14%) | 10 (43%) | 17 (57%) | 2 (6%) | 0 (0%) | 7 (29%) | 1 (4%) | 2 (7%) |
| Yes | 10 (62%) | 8 (73%) | 19 (86%) | 13 (57%) | 13 (43%) | 32 (94%) | 14 (100%) | 17 (71%) | 22 (96%) | 25 (93%) |
| **Comparability of eligibility criteria** | | | | | | | | | | |
| Narrower | 6 (38%) | 2 (18%) | 18 (82%) | 18 (78%) | 28 (93%) | 22 (65%) | 0 (0%) | 4 (17%) | 4 (17%) | 20 (74%) |
| Comparable | 4 (25%) | 0 (0%) | 2 (9%) | 3 (13%) | 2 (7%) | 5 (15%) | 0 (0%) | 1 (4%) | 3 (13%) | 7 (26%) |
| Mixture | 5 (31%) | 9 (82%) | 0 (0%) | 2 (9%) | 0 (0%) | 3 (9%) | 0 (0%) | 16 (67%) | 11 (48%) | 0 (0%) |
| Broader | 1 (6%) | 0 (0%) | 2 (9%) | 0 (0%) | 0 (0%) | 2 (6%) | 0 (0%) | 3 (12%) | 5 (22%) | 0 (0%) |
| Non-overlapping | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 14 (100%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Unclear | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 2 (6%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| **Comparability of age eligibility criteria** | | | | | | | | | | |
| Narrower | 1 (6%) | 2 (18%) | 0 (0%) | 17 (74%) | 9 (30%) | 13 (38%) | 10 (71%) | 2 (8%) | 1 (4%) | 17 (63%) |
| Comparable | 15 (94%) | 0 (0%) | 22 (100%) | 0 (0%) | 1 (3%) | 21 (62%) | 4 (29%) | 15 (62%) | 4 (17%) | 4 (15%) |
| Mixture | 0 (0%) | 0 (0%) | 0 (0%) | 6 (26%) | 16 (53%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Broader | 0 (0%) | 9 (82%) | 0 (0%) | 0 (0%) | 4 (13%) | 0 (0%) | 0 (0%) | 7 (29%) | 18 (78%) | 6 (22%) |
| **Setting** | | | | | | | | | | |
| Primary care | 3 (19%) | 0 (0%) | 0 (0%) | 7 (30%) | 3 (10%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Secondary care | 12 (75%) | 9 (82%) | 16 (84%) | 0 (0%) | 5 (17%) | 18 (82%) | 6 (75%) | 18 (90%) | 12 (86%) | 4 (40%) |
| Tertiary care | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |

| | ABCD2 | ESRS | EuroSCORE | Framingham | FRAX | ISS | MELD | PSI | RCRI | SAPS 3 |
|---|---|---|---|---|---|---|---|---|---|---|
| Population based | 0 (0%) | 1 (9%) | 0 (0%) | 15 (65%) | 17 (59%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Screening | 0 (0%) | 0 (0%) | 0 (0%) | 1 (4%) | 1 (3%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Mixed | 1 (6%) | 1 (9%) | 2 (11%) | 0 (0%) | 0 (0%) | 4 (18%) | 1 (12%) | 2 (10%) | 2 (14%) | 2 (20%) |
| Unclear | 0 (0%) | 0 (0%) | 1 (5%) | 0 (0%) | 3 (10%) | 0 (0%) | 1 (12%) | 0 (0%) | 0 (0%) | 4 (40%) |
| **Comparability of setting** | | | | | | | | | | |
| Comparable | 1 (6%) | 0 (0%) | 16 (73%) | 15 (65%) | 17 (57%) | 18 (53%) | 6 (43%) | 18 (75%) | 9 (39%) | 4 (15%) |
| Narrower | 15 (94%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Broader | 0 (0%) | 0 (0%) | 2 (9%) | 0 (0%) | 0 (0%) | 4 (12%) | 1 (7%) | 2 (8%) | 2 (9%) | 2 (7%) |
| Non-overlapping | 0 (0%) | 11 (100%) | 3 (14%) | 8 (35%) | 10 (33%) | 12 (35%) | 6 (43%) | 4 (17%) | 12 (52%) | 17 (63%) |
| Unclear | 0 (0%) | 0 (0%) | 1 (5%) | 0 (0%) | 3 (10%) | 0 (0%) | 1 (7%) | 0 (0%) | 0 (0%) | 4 (15%) |
| **Continent** | | | | | | | | | | |
| Africa | 0 (0%) | 0 (0%) | 1 (5%) | 0 (0%) | 0 (0%) | 1 (3%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Asia | 0 (0%) | 2 (18%) | 4 (18%) | 2 (9%) | 4 (13%) | 7 (21%) | 1 (7%) | 3 (12%) | 2 (9%) | 5 (19%) |
| Australia | 0 (0%) | 0 (0%) | 1 (5%) | 0 (0%) | 5 (17%) | 2 (6%) | 0 (0%) | 3 (12%) | 1 (4%) | 1 (4%) |
| Europe | 8 (50%) | 7 (64%) | 10 (45%) | 10 (43%) | 9 (30%) | 7 (21%) | 7 (50%) | 11 (46%) | 13 (57%) | 9 (33%) |
| North America | 8 (50%) | 1 (9%) | 5 (23%) | 11 (48%) | 11 (37%) | 17 (50%) | 3 (21%) | 6 (25%) | 6 (26%) | 3 (11%) |
| South America | 0 (0%) | 0 (0%) | 1 (5%) | 0 (0%) | 0 (0%) | 0 (0%) | 3 (21%) | 0 (0%) | 0 (0%) | 9 (33%) |
| Combination | 0 (0%) | 1 (9%) | 0 (0%) | 0 (0%) | 1 (3%) | 0 (0%) | 0 (0%) | 1 (4%) | 1 (4%) | 0 (0%) |
| **Comparability of continent** | | | | | | | | | | |
| Comparable | 0 (0%) | 0 (0%) | 10 (45%) | 11 (48%) | 0 (0%) | 17 (50%) | 3 (21%) | 6 (25%) | 6 (26%) | 0 (0%) |
| Narrower | 16 (100%) | 8 (73%) | 0 (0%) | 0 (0%) | 30 (100%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 27 (100%) |
| Broader | 0 (0%) | 1 (9%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (4%) | 1 (4%) | 0 (0%) |
| Non-overlapping | 0 (0%) | 2 (18%) | 12 (55%) | 12 (52%) | 0 (0%) | 17 (50%) | 11 (79%) | 17 (71%) | 16 (70%) | 0 (0%) |
| **Number of centres** | | | | | | | | | | |
| Single | 9 (56%) | 4 (36%) | 12 (55%) | 3 (13%) | 5 (17%) | 18 (53%) | 12 (86%) | 14 (58%) | 17 (74%) | 13 (48%) |
| Multiple | 6 (38%) | 7 (64%) | 9 (41%) | 6 (26%) | 9 (30%) | 15 (44%) | 2 (14%) | 10 (42%) | 6 (26%) | 14 (52%) |
| Population based | 1 (6%) | 0 (0%) | 0 (0%) | 12 (52%) | 15 (50%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Unclear | 0 (0%) | 0 (0%) | 1 (5%) | 2 (9%) | 1 (3%) | 1 (3%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| **Changes made to predictors** | | | | | | | | | | |
| No | 16 (100%) | 10 (91%) | 13 (59%) | 23 (100%) | 20 (67%) | 21 (62%) | 12 (86%) | 13 (54%) | 17 (74%) | 26 (96%) |
| Yes | 0 (0%) | 0 (0%) | 5 (23%) | 0 (0%) | 10 (33%) | 0 (0%) | 0 (0%) | 5 (21%) | 2 (9%) | 0 (0%) |
| Unclear | 0 (0%) | 1 (9%) | 4 (18%) | 0 (0%) | 0 (0%) | 13 (38%) | 2 (14%) | 6 (25%) | 4 (17%) | 1 (4%) |
| **Comparability of outcome definition** | | | | | | | | | | |

| | ABCD2 | ESRS | EuroSCORE | Framingham | FRAX | ISS | MELD | PSI | RCRI | SAPS 3 |
|---|---|---|---|---|---|---|---|---|---|---|
| No | 8 (50%) | 3 (27%) | 9 (41%) | 13 (57%) | 16 (53%) | 7 (21%) | 14 (100%) | 5 (21%) | 5 (22%) | 24 (89%) |
| Yes | 3 (19%) | 8 (73%) | 11 (50%) | 4 (17%) | 13 (43%) | 19 (56%) | 0 (0%) | 18 (75%) | 18 (78%) | 3 (11%) |
| Unclear | 5 (31%) | 0 (0%) | 2 (9%) | 6 (26%) | 1 (3%) | 8 (24%) | 0 (0%) | 1 (4%) | 0 (0%) | 0 (0%) |
| **Outcome measurement method** | | | | | | | | | | |
| Self-reported | 3 (19%) | 8 (73%) | 2 (9%) | 15 (65%) | 18 (60%) | 0 (0%) | 1 (7%) | 6 (25%) | 1 (4%) | 2 (7%) |
| Clinician | 6 (38%) | 1 (9%) | 2 (9%) | 2 (9%) | 1 (3%) | 4 (12%) | 0 (0%) | 2 (8%) | 10 (43%) | 6 (22%) |
| Registry | 3 (19%) | 2 (18%) | 5 (23%) | 4 (17%) | 8 (27%) | 13 (38%) | 3 (21%) | 9 (38%) | 6 (26%) | 9 (33%) |
| Unclear | 4 (25%) | 0 (0%) | 13 (59%) | 2 (9%) | 3 (10%) | 17 (50%) | 10 (71%) | 7 (29%) | 6 (26%) | 10 (37%) |
| **Similar outcome measurement for all patients** | | | | | | | | | | |
| Yes | 9 (56%) | 3 (27%) | 11 (50%) | 6 (26%) | 14 (47%) | 18 (53%) | 2 (14%) | 12 (50%) | 13 (57%) | 16 (59%) |
| No | 1 (6%) | 5 (45%) | 1 (5%) | 16 (70%) | 3 (10%) | 3 (9%) | 0 (0%) | 3 (12%) | 4 (17%) | 1 (4%) |
| Unclear | 6 (38%) | 3 (27%) | 10 (45%) | 1 (4%) | 13 (43%) | 13 (38%) | 12 (86%) | 9 (38%) | 6 (26%) | 10 (37%) |
| **Method for handling of missing data** | | | | | | | | | | |
| Complete case analysis | 4 (25%) | 8 (73%) | 3 (14%) | 11 (48%) | 19 (63%) | 18 (53%) | 7 (50%) | 1 (4%) | 5 (22%) | 8 (30%) |
| Mean/median imputation | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (4%) |
| Multiple imputation | 1 (6%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| NA | 3 (19%) | 1 (9%) | 1 (5%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (4%) | 1 (4%) | 2 (7%) |
| Other | 0 (0%) | 0 (0%) | 1 (5%) | 0 (0%) | 2 (7%) | 0 (0%) | 1 (7%) | 8 (33%) | 0 (0%) | 7 (26%) |
| Unclear | 8 (50%) | 2 (18%) | 17 (77%) | 12 (52%) | 9 (30%) | 16 (47%) | 6 (43%) | 14 (58%) | 17 (74%) | 9 (33%) |
| **Handling of missing data** | | | | | | | | | | |
| Appropriate or <5% missing | 8 (50%) | 4 (36%) | 2 (9%) | 3 (13%) | 4 (13%) | 6 (18%) | 4 (29%) | 4 (17%) | 7 (30%) | 6 (22%) |
| Inappropriate | 1 (6%) | 5 (45%) | 3 (14%) | 8 (35%) | 17 (57%) | 13 (38%) | 4 (29%) | 8 (33%) | 0 (0%) | 12 (44%) |
| Unclear | 7 (44%) | 2 (18%) | 17 (77%) | 12 (52%) | 9 (30%) | 15 (44%) | 6 (43%) | 12 (50%) | 16 (70%) | 9 (33%) |

Continuous variables

| | ABCD2 | ESRS | EuroSCORE | Framingham | FRAX | ISS | MELD | PSI | RCRI | SAPSIII |
|---|---|---|---|---|---|---|---|---|---|---|
| Year start recruitment | 2002 (2000-2003) NR=0 | 2007 (2004-2007) NR=0 | 1998 (1995-2001) NR=1 | 1989 (1983-1994) NR=0 | 1994 (1990-1998) NR=3 | 1996 (1993-1998) NR=1 | 2000 (1998-2004) NR=0 | 2000 (1998-2002) NR=0 | 2000 (1994-2002) NR=4 | 2006 (2006-2007) NR=0 |
| Year end recruitment | 2005 (2003-2007) NR=0 | 2008 (2006-2008) NR=0 | 2002 (1999-2005) NR=1 | 1993 (1988-1998) NR=0 | 1997 (1993-2006) NR=8 | 2000 (1996-2003) NR=2 | 2006 (2004-2007) NR=0 | 2002 (2000-2003) NR=0 | 2002 (2000-2005) NR=4 | 2007 (2006-2009) NR=0 |
| Percentage missings | 0.95 (0.00-5.00) NR=7 | 5.12 (1.99-17.80) NR=2 | 6.40 (1.50-11.83) NR=18 | 4.90 (2.70-9.80) NR=18 | 30.25 (2.75-33.80) NR=16 | 9.05 (2.40-14.65) NR=20 | 4.05 (2.73-10.93) NR=8 | 0.52 (0.07-9.26) NR=18 | 1.00 (0.09-1.91) NR=16 | 5.85 (0.52-18.93) NR=15 |
| Number of participants | 304 (204-691) NR=0 | 1257 (712-2594) NR=0 | 1730 (873-4518) NR=2 | 2399 (928-4609) NR=0 | 2210 (889-6586) NR=0 | 2590 (960-20713) NR=0 | 418 (118-483) NR=0 | 730 (326-970) NR=1 | 496 (180-1480) NR=0 | 864 (485-1856) NR=0 |
| Number of events | 9 (3-18) NR=0 | 92 (60-134) NR=0 | 36 (13-87) NR=2 | 92 (72-160) NR=1 | 250 (86-581) NR=0 | 256 (113-1660) NR=2 | 49 (22-112) NR=0 | 54 (28-111) NR=1 | 31 (14-76) NR=0 | 180 (124-311) NR=1 |
| Age mean | 67.4 (64.1-70.0) NR=5 | 68.3 (67.1-71.5) NR=3 | 63.9 (62.5-65.2) NR=2 | 54.6 (50.9-58.3) NR=2 | 66.8 (63.0-71.3) NR=1 | 38.1 (32.4-41.3) NR=10 | 51.8 (49.1-53.0) NR=0 | 66.2 (64.0-69.3) NR=2 | 67.8 (66.0-71.9) NR=2 | 62.2 (60.8-64.8) NR=1 |
| Age sd | 13.8 (13.0-14.9) NR=5 | 12.4 (12.0-13.0) NR=1 | 9.3 (9.0-10.6) NR=8 | 7.3 (4.1-9.4) NR=0 | 8.3 (5.9-9.8) NR=0 | 20.9 (18.1-24.8) NR=2 | 10.0 (9.6-12.0) NR=1 | 17.8 (17.0-20.1) NR=3 | 10.0 (8.8-12.5) NR=4 | 17.0 (15.4-19.0) NR=3 |
| Gender percentage men | 47 (45-53) NR=2 | 57 (55-59) NR=1 | 77 (71-79) NR=1 | 100 (100-100) NR=0 | 0 (0-0) NR=0 | 71 (64-75) NR=11 | 68 (63-69) NR=0 | 57 (53-64) NR=1 | 67 (52-76) NR=0 | 59 (55-64) NR=0 |

Values represent median (IQR), number of missing values.

Figure S1: Associations between categorical variables and c-statistic

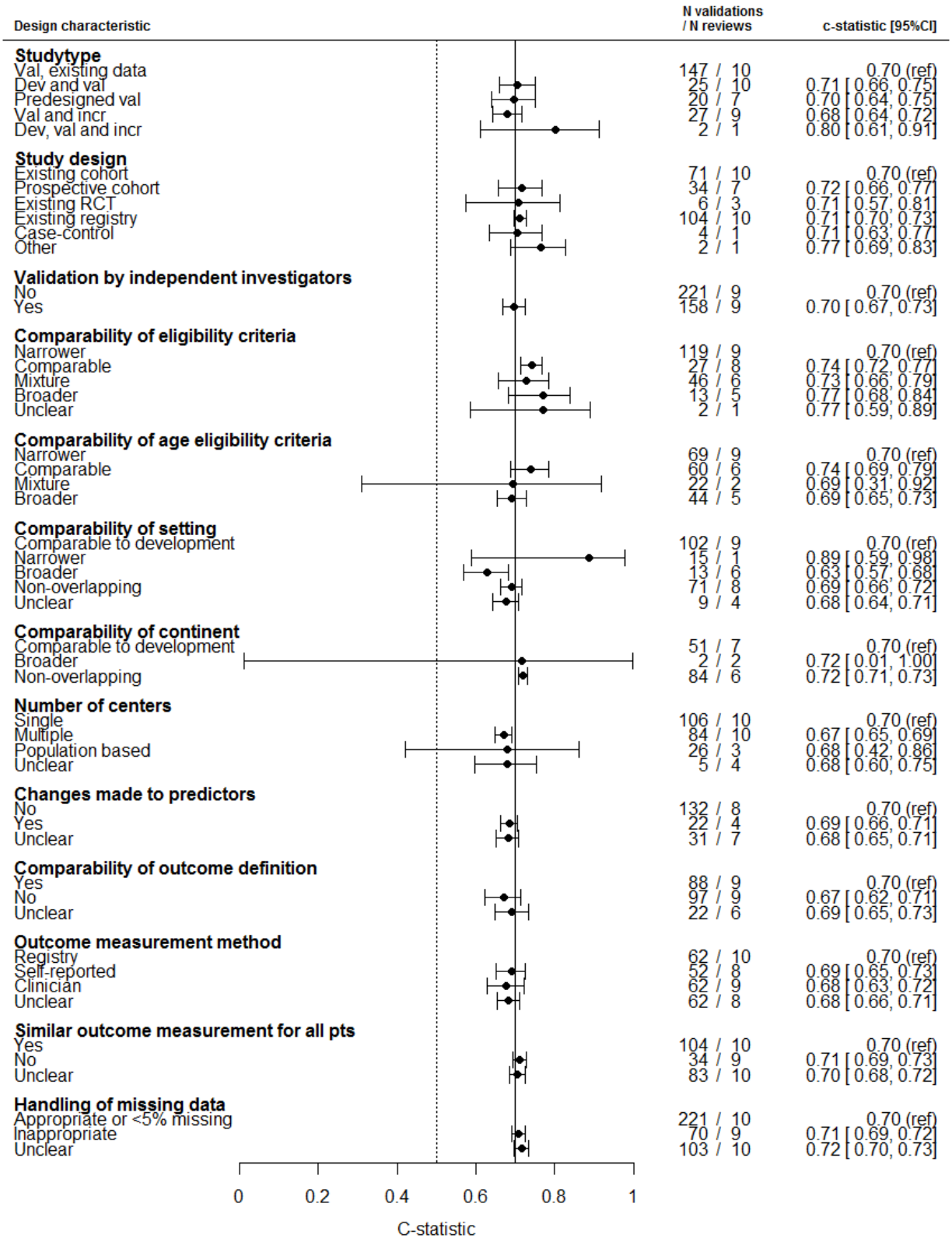
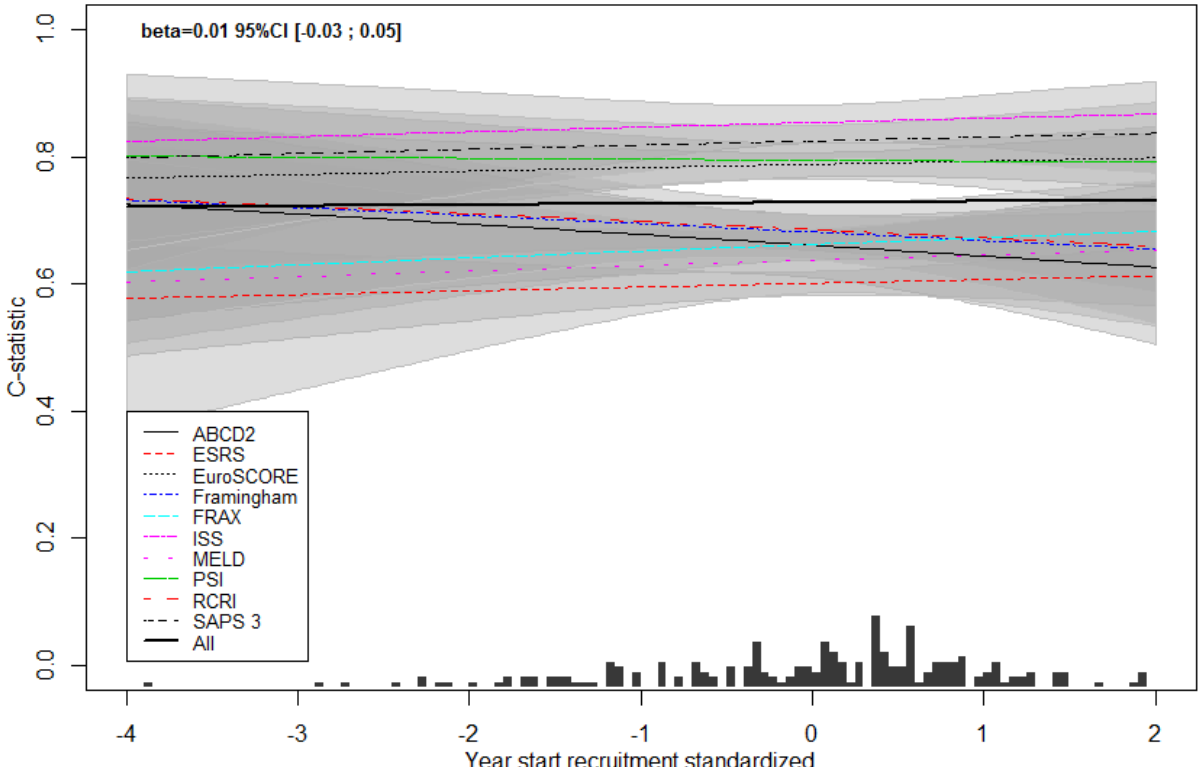
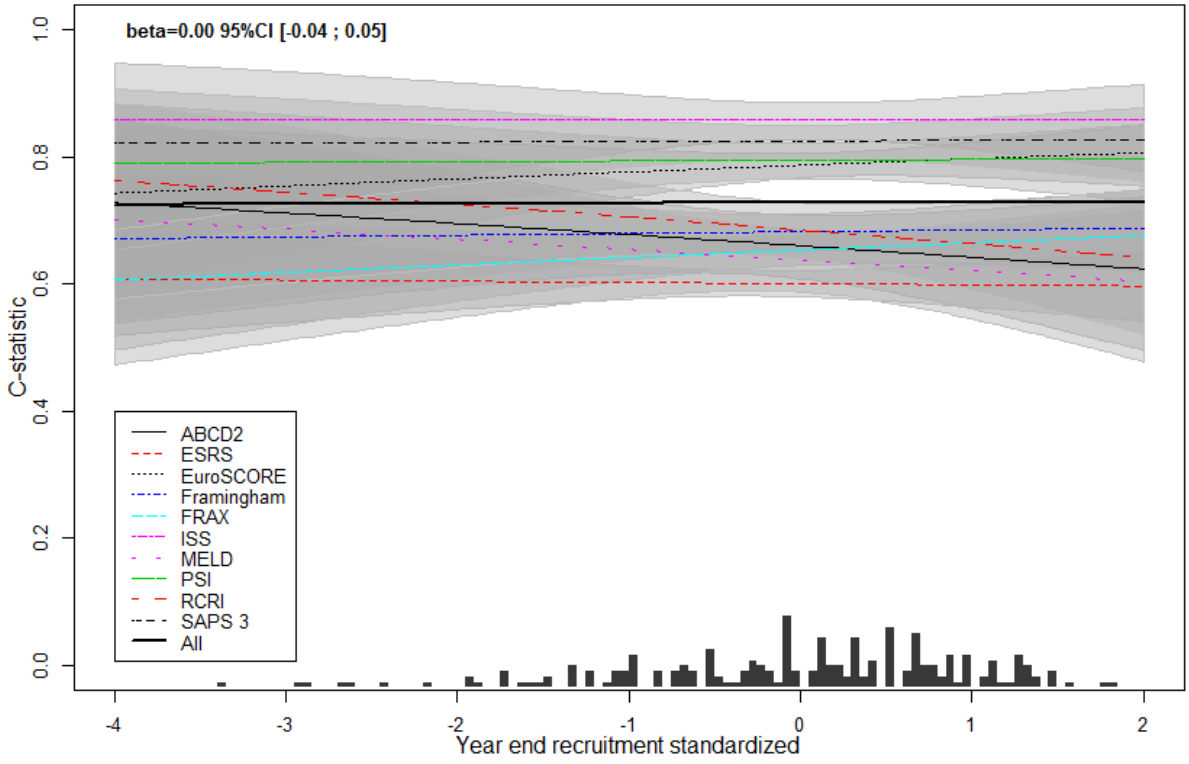
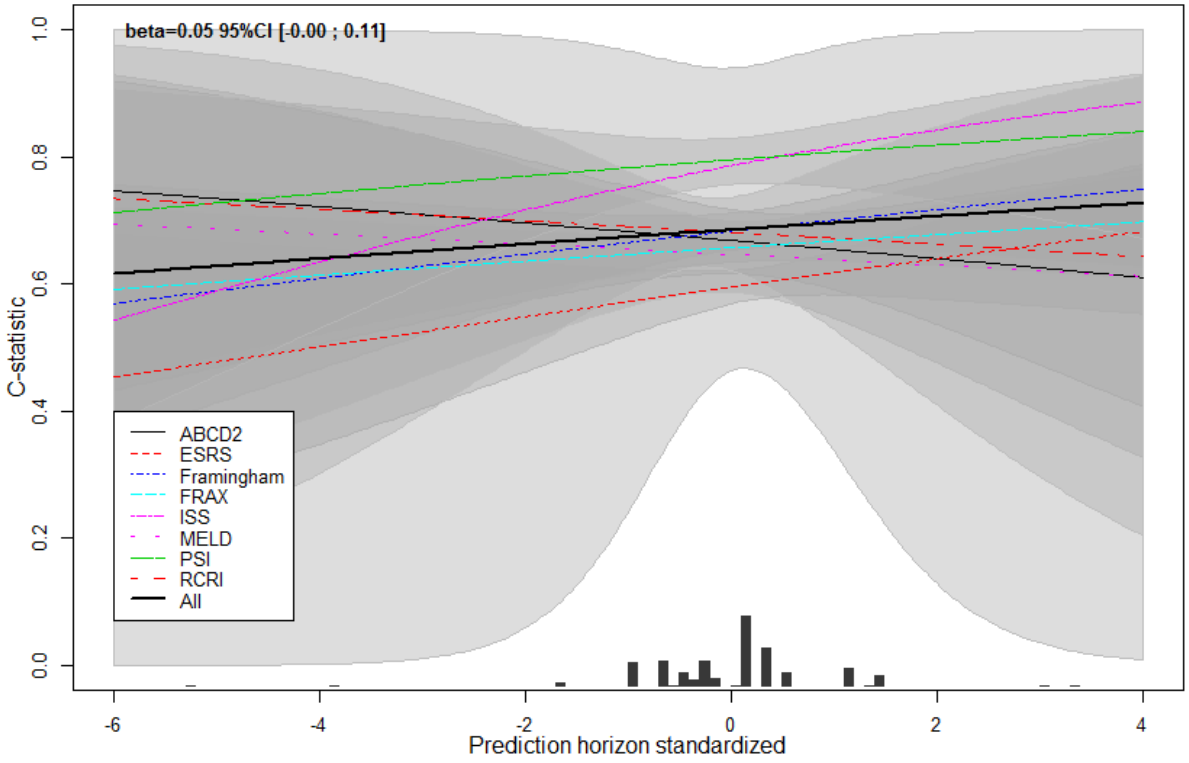
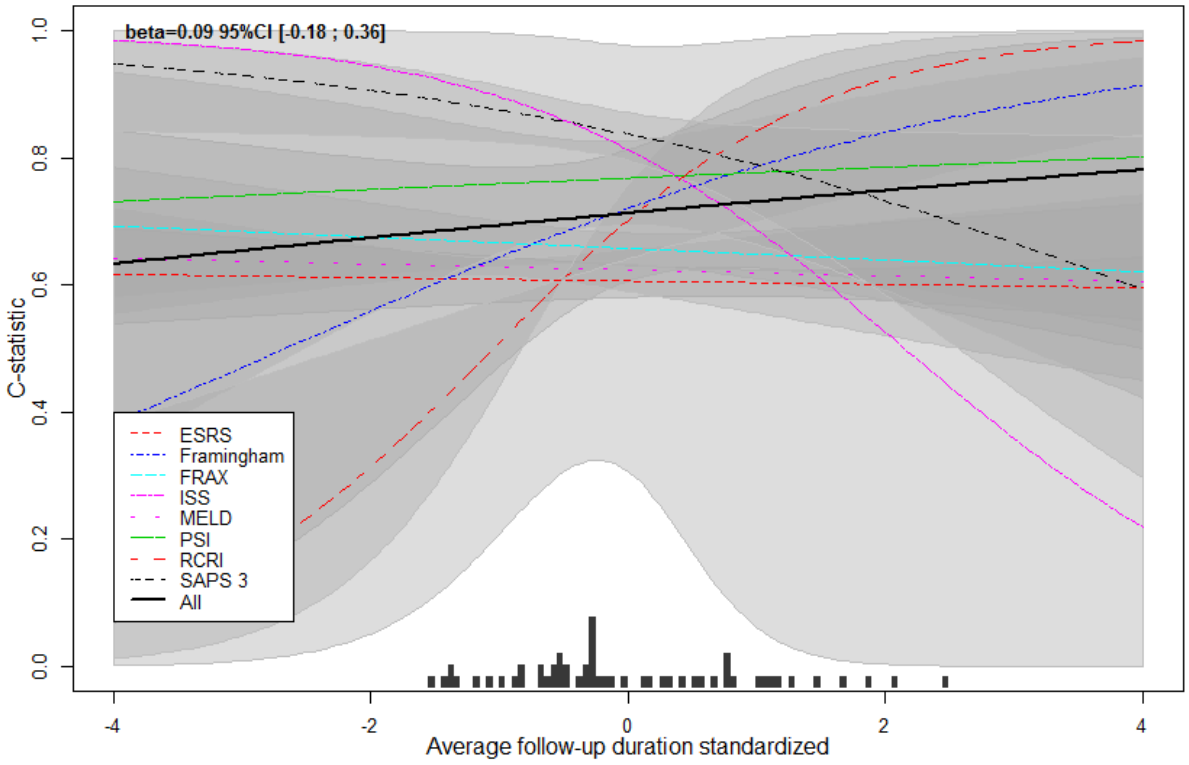


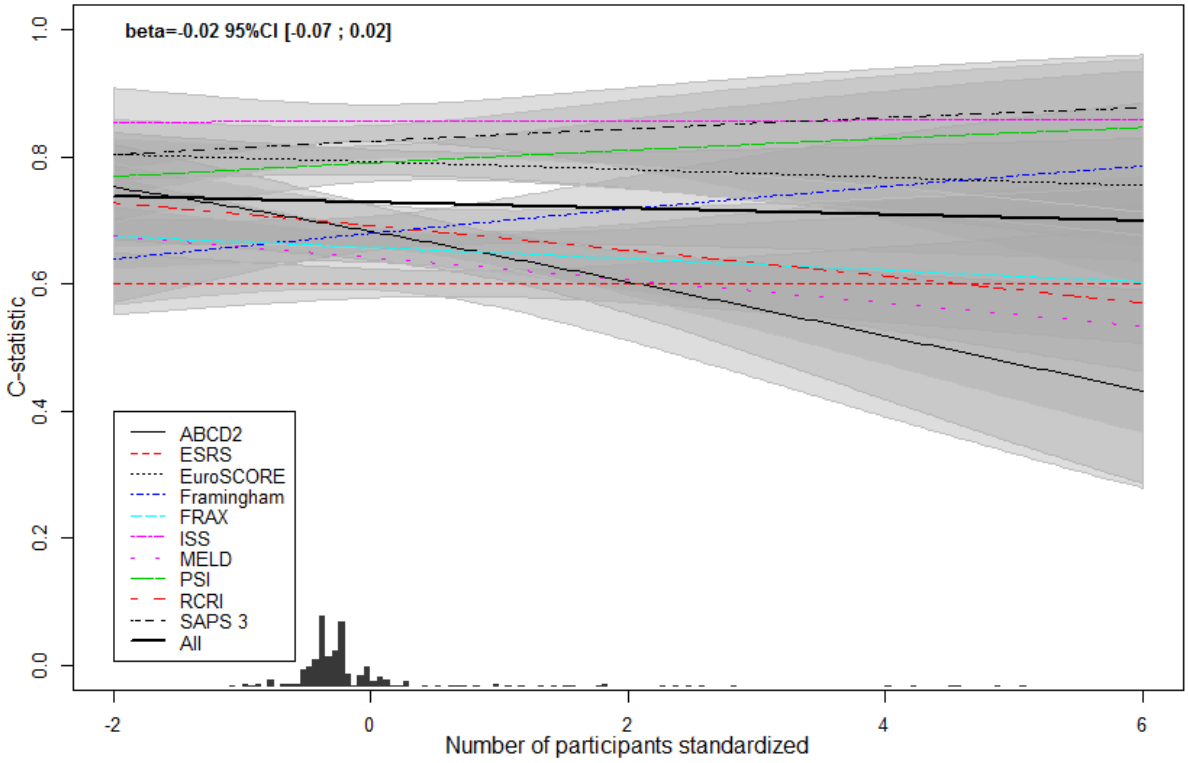
Figure S2: Associations between continuous variables and c-statistic

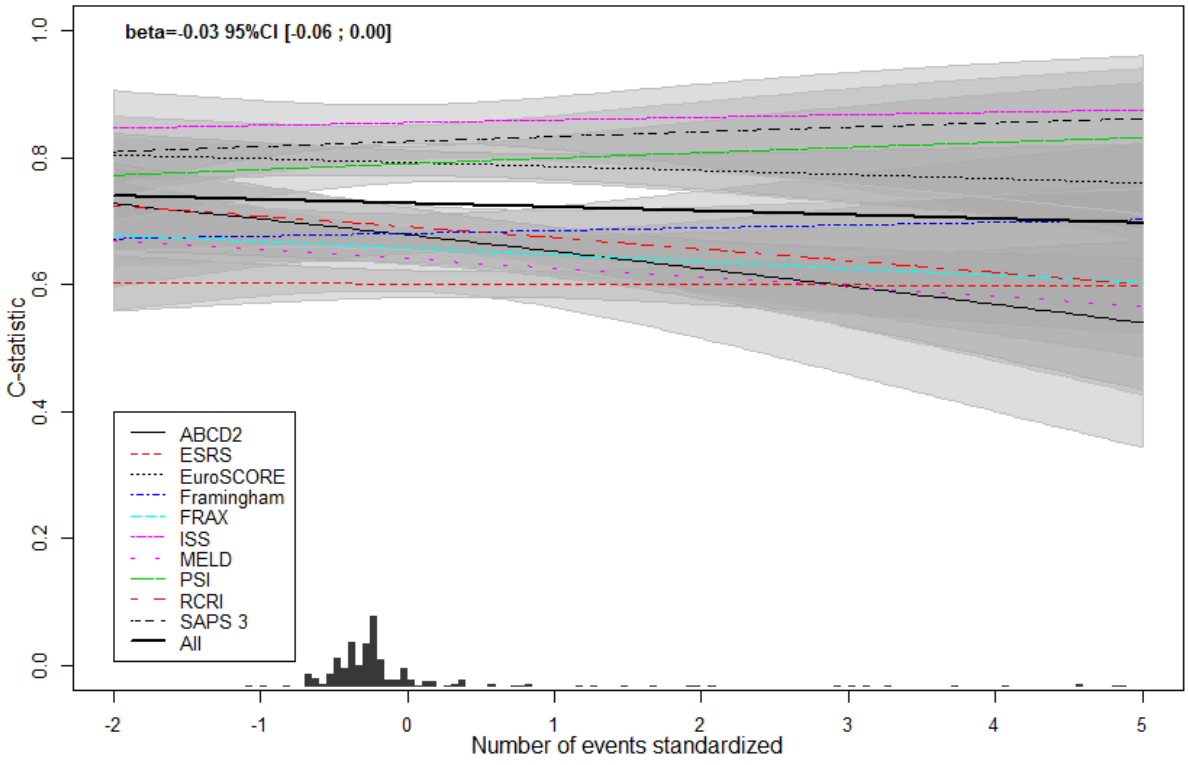


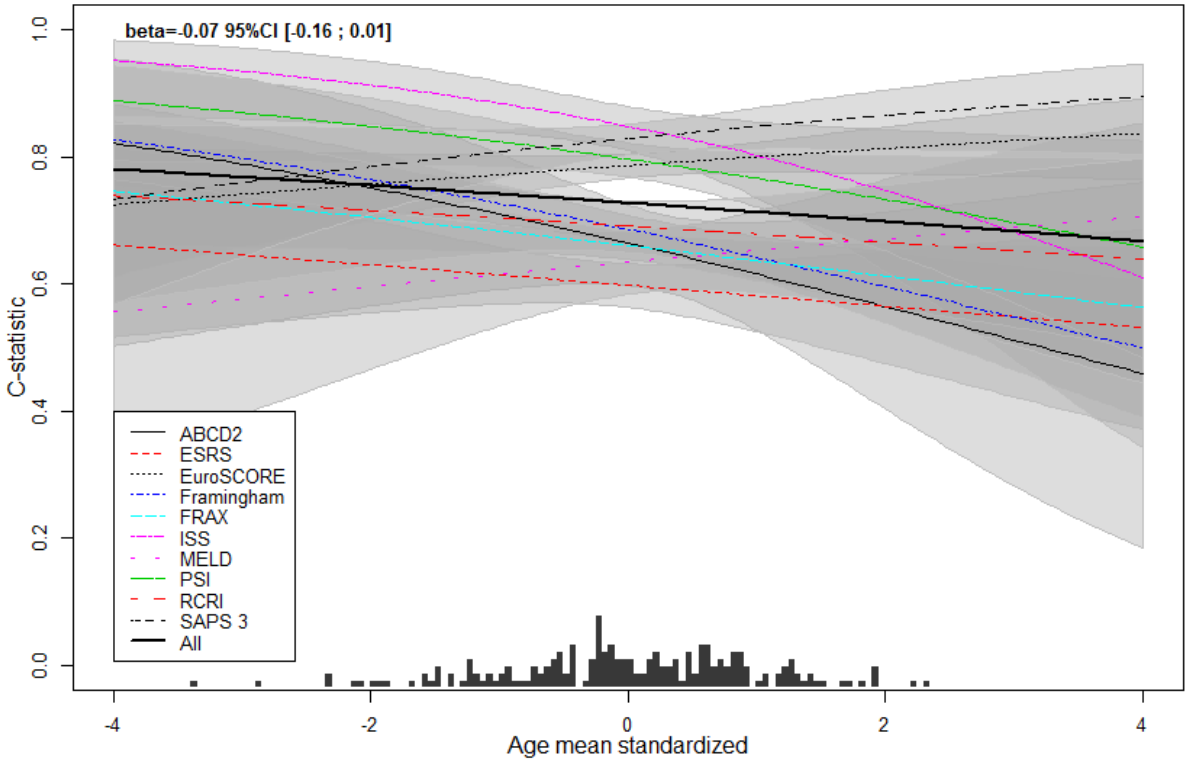



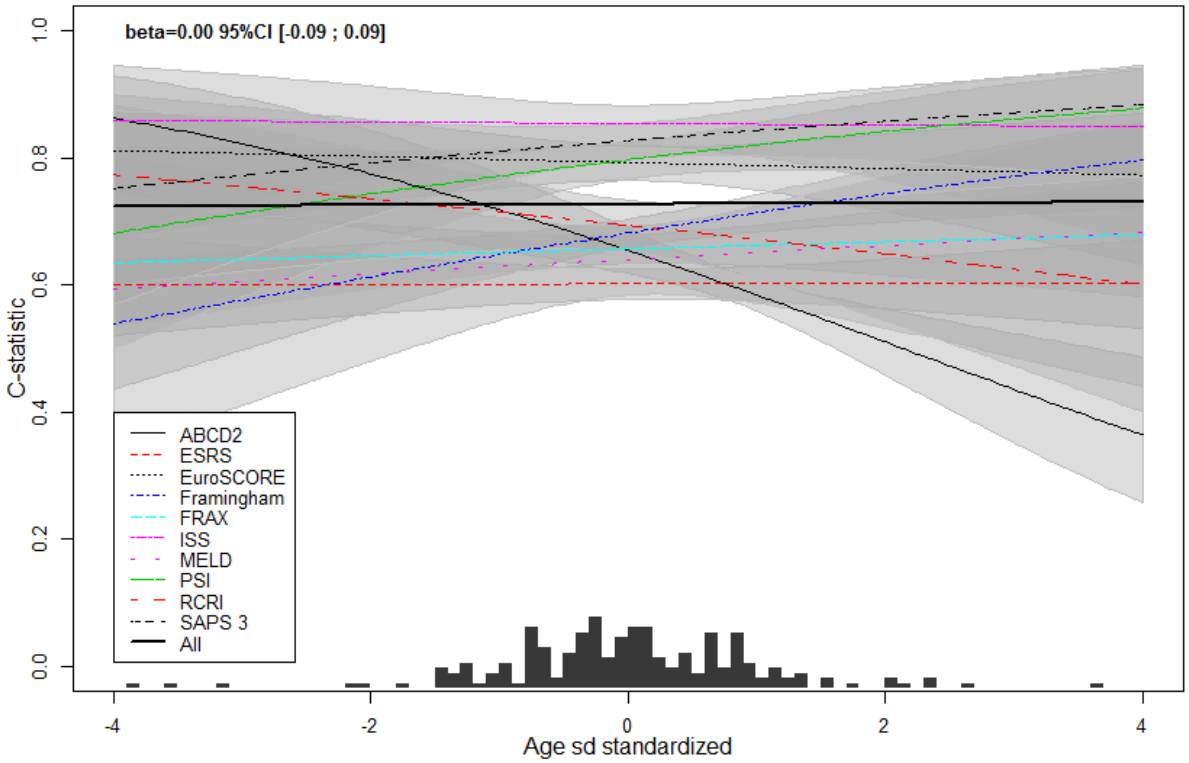

beta=0.05 95%CI [-0.00 ; 0.11]
C-statistic
0.0
0.2
0.4
0.6
0.8
1.0
ABCD2
ESRS
Framingham
FRAX
ISS
MELD
PSI
RCRI
All
-6
-4
-2
0
2
4
Prediction horizon standardized


beta=0.09 95%CI [-0.18 ; 0.36]
C-statistic
0.0
0.2
0.4
0.6
0.8
1.0
ESRS
Framingham
FRAX
ISS
MELD
PSI
RCRI
SAPS 3
All
-4
-2
0
2
4
Average follow-up duration standardized


beta=-0.02 95%CI [-0.07 ; 0.02]
C-statistic
0.0
0.2
0.4
0.6
0.8
1.0
-2
0
2
4
6
Number of participants standardized
ABCD2
ESRS
EuroSCORE
Framingham
FRAX
ISS
MELD
PSI
RCRI
SAPS 3
All


beta=-0.03 95%CI [-0.06 ; 0.00]
C-statistic
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
3
4
5
Number of events standardized
ABCD2
ESRS
EuroSCORE
Framingham
FRAX
ISS
MELD
PSI
RCRI
SAPS 3
All


beta=-0.07 95%CI [-0.16 ; 0.01]
C-statistic
0.0
0.2
0.4
0.6
0.8
1.0
ABCD2
ESRS
EuroSCORE
Framingham
FRAX
ISS
MELD
PSI
RCRI
SAPS 3
All
-4
-2
0
2
4
Age mean standardized


beta=0.00 95%CI [-0.09 ; 0.09]
C-statistic
0.0
0.2
0.4
0.6
0.8
1.0
ABCD2
ESRS
EuroSCORE
Framingham
FRAX
ISS
MELD
PSI
RCRI
SAPS 3
All
-4
-2
0
2
4
Age sd standardized

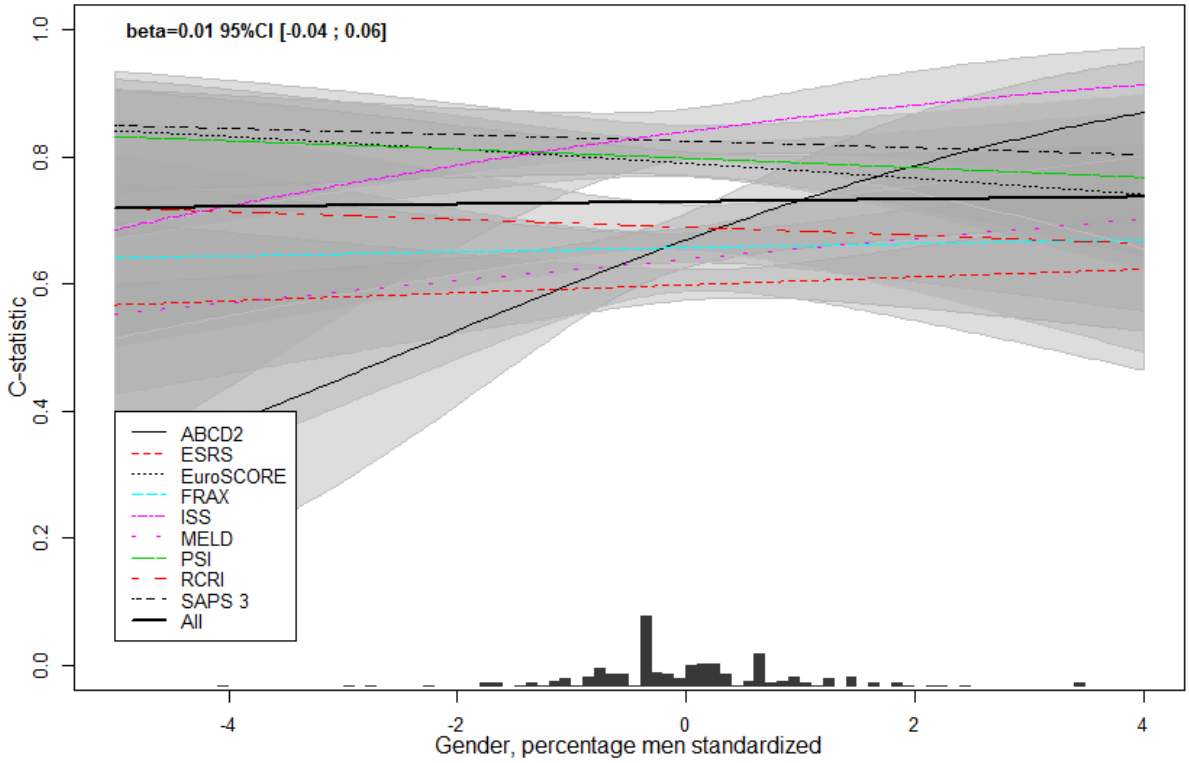
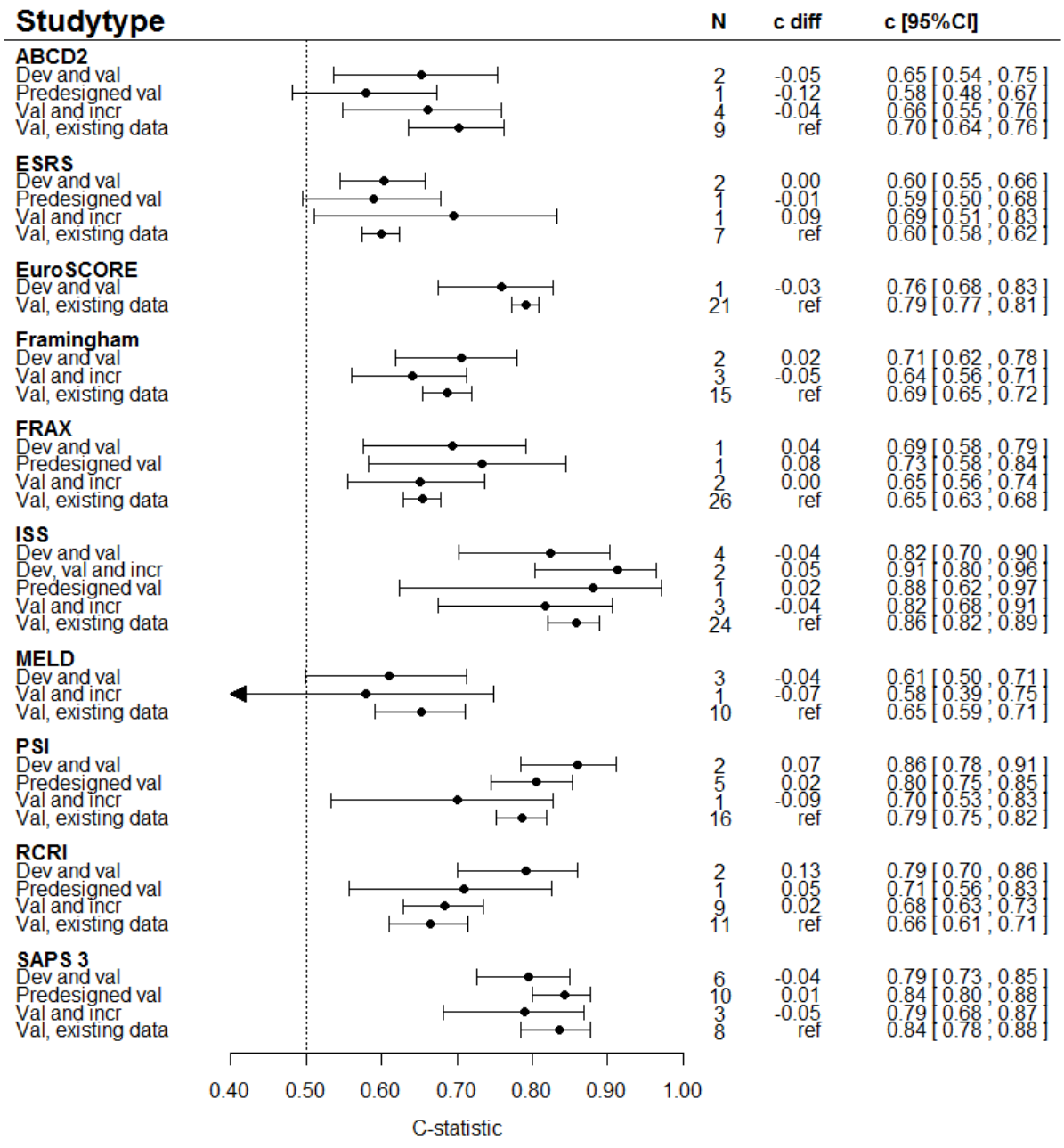
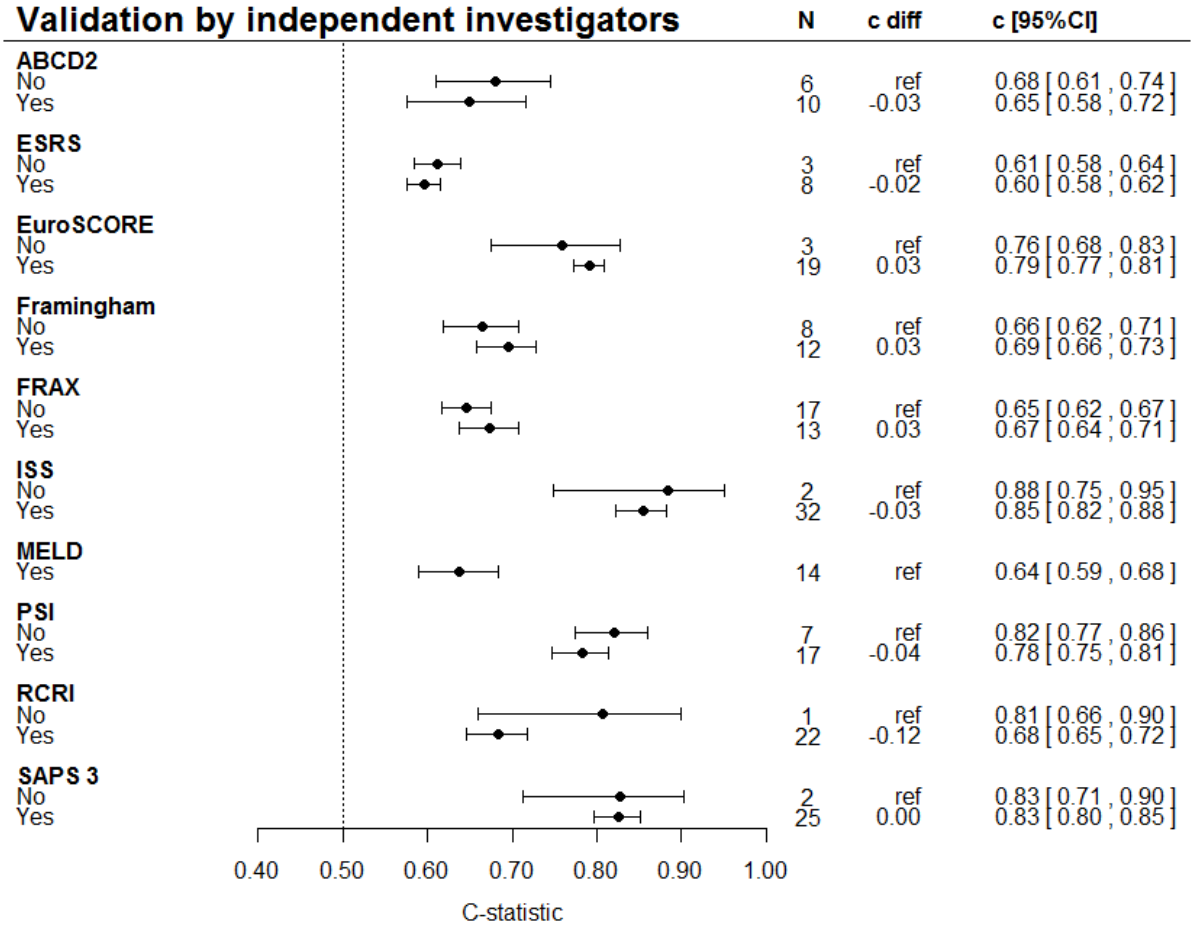
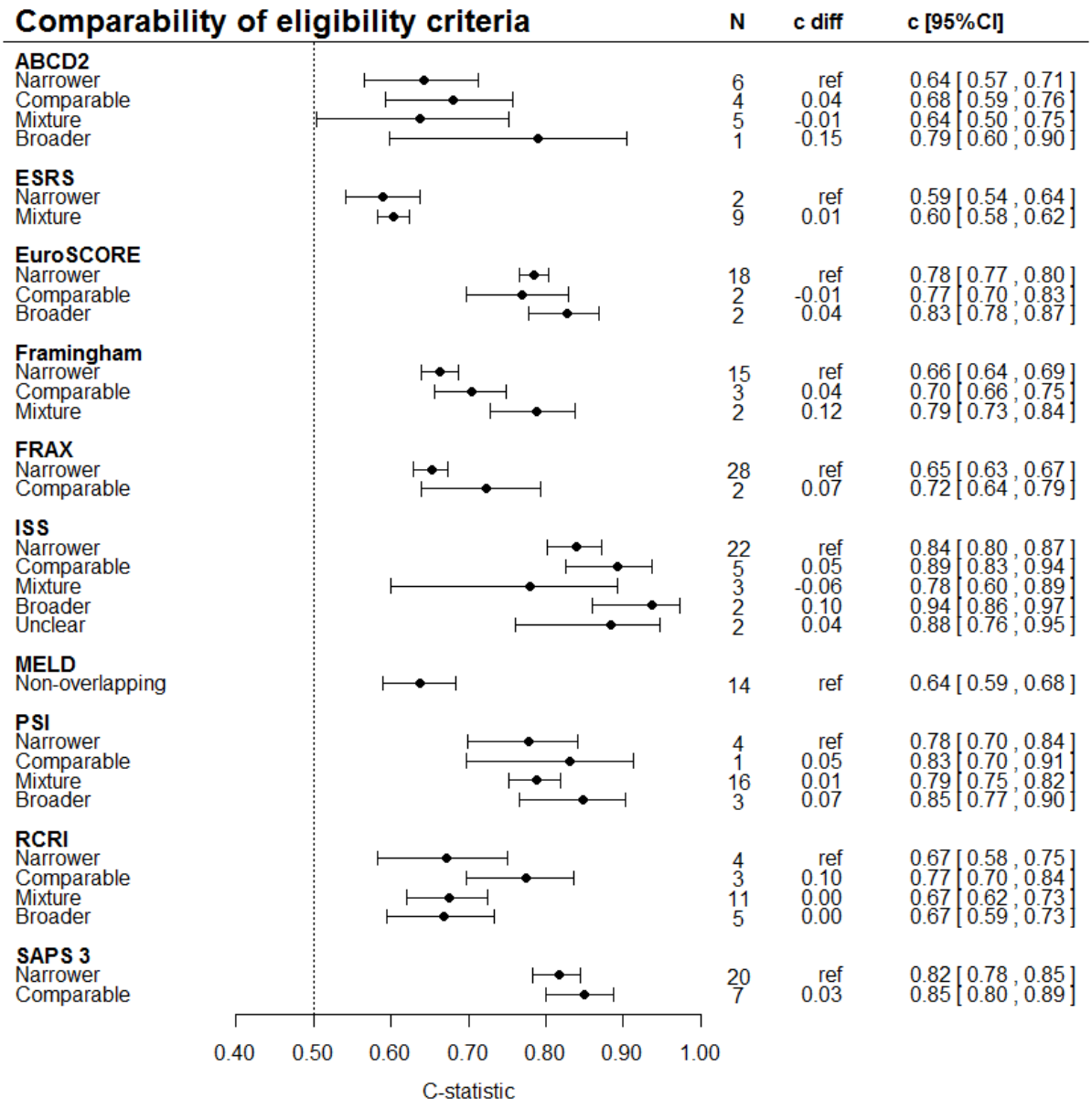
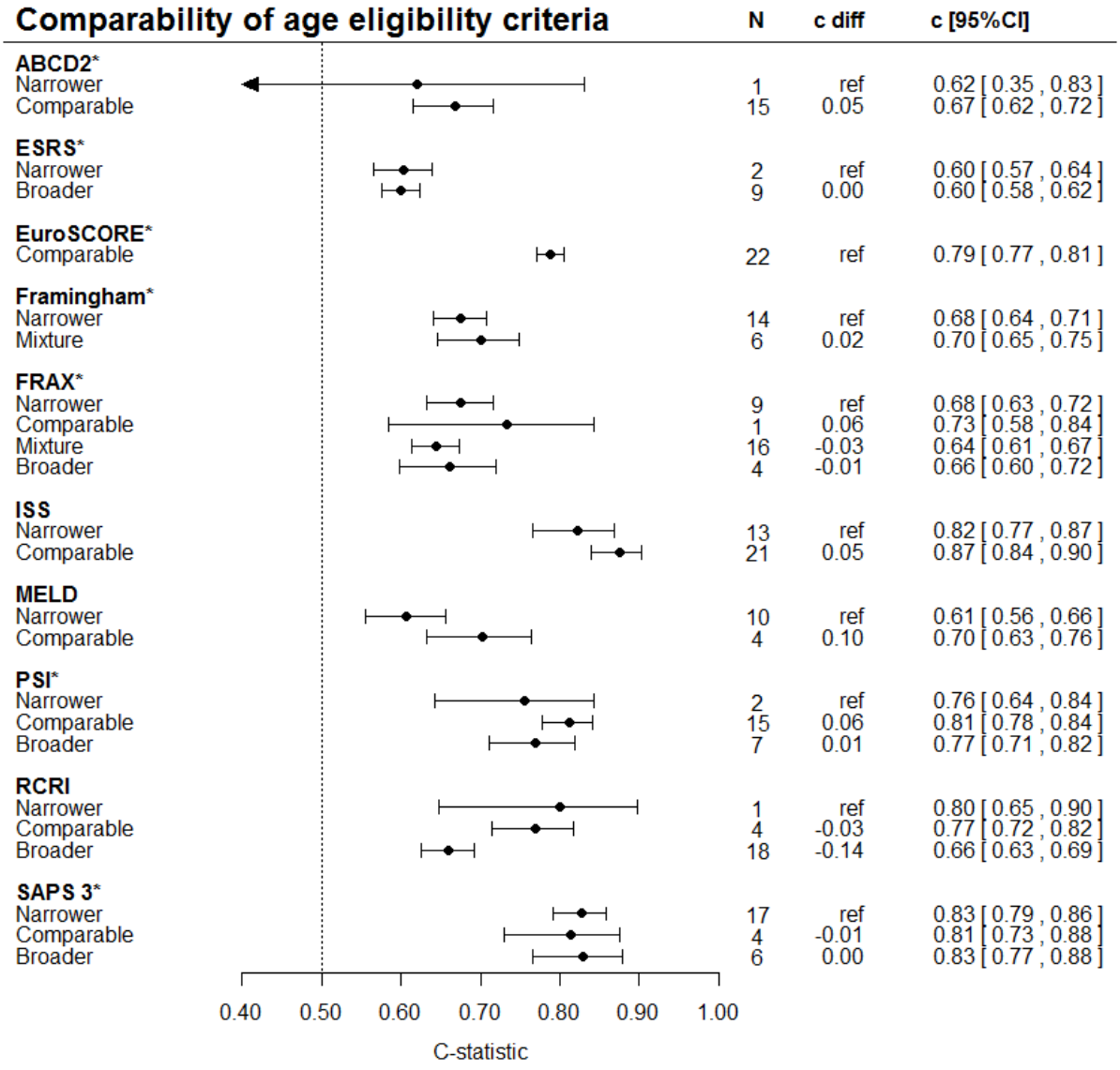

beta=0.01 95%CI [-0.04 ; 0.06]
C-statistic
0.0
0.2
0.4
0.6
0.8
1.0
ABCD2
ESRS
EuroSCORE
FRAX
ISS
MELD
PSI
RCRI
SAPS 3
All
-4
-2
0
2
4
Gender, percentage men standardized

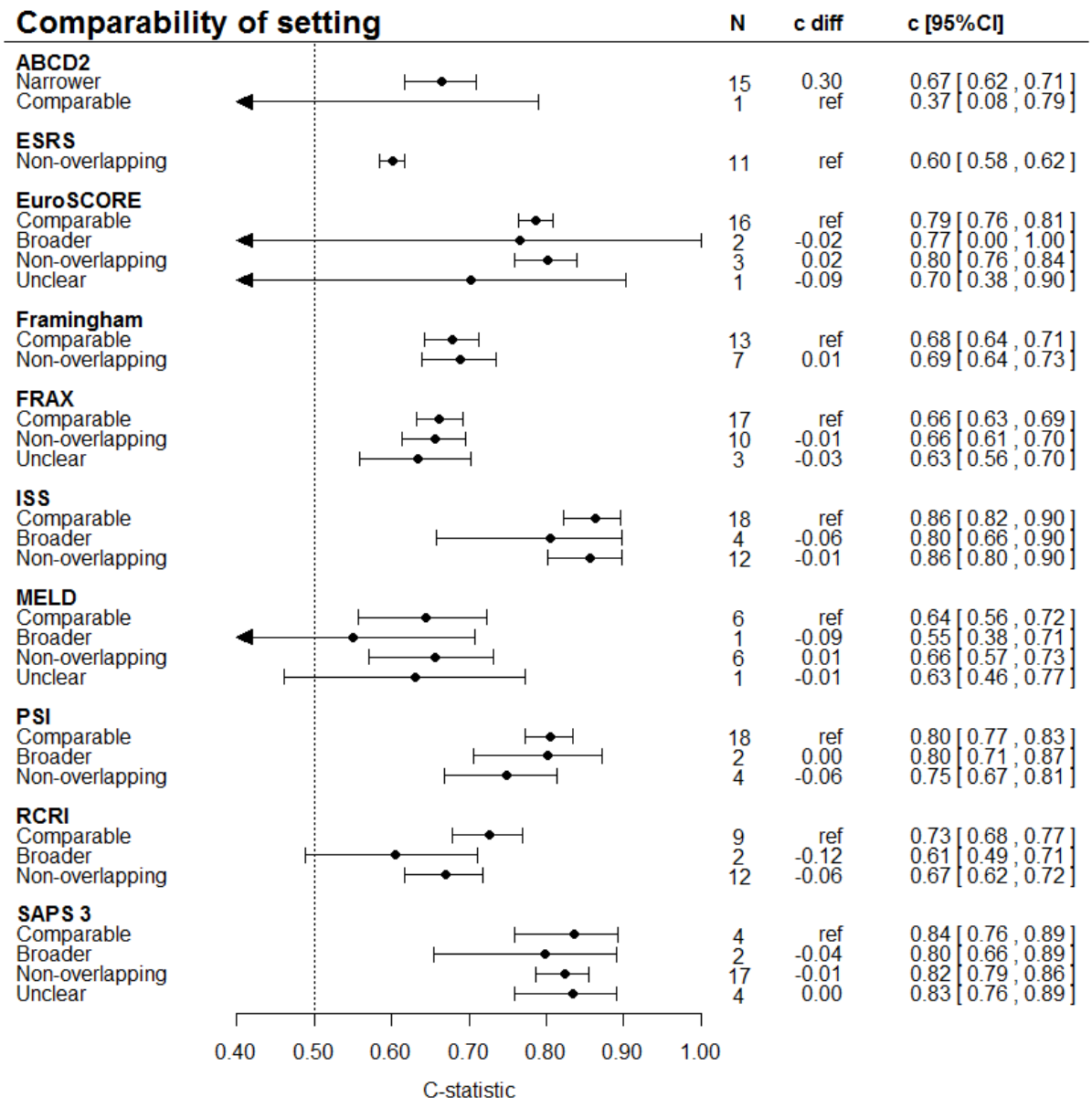
Figure S3: C-statistic in categories of study characteristics within each systematic review

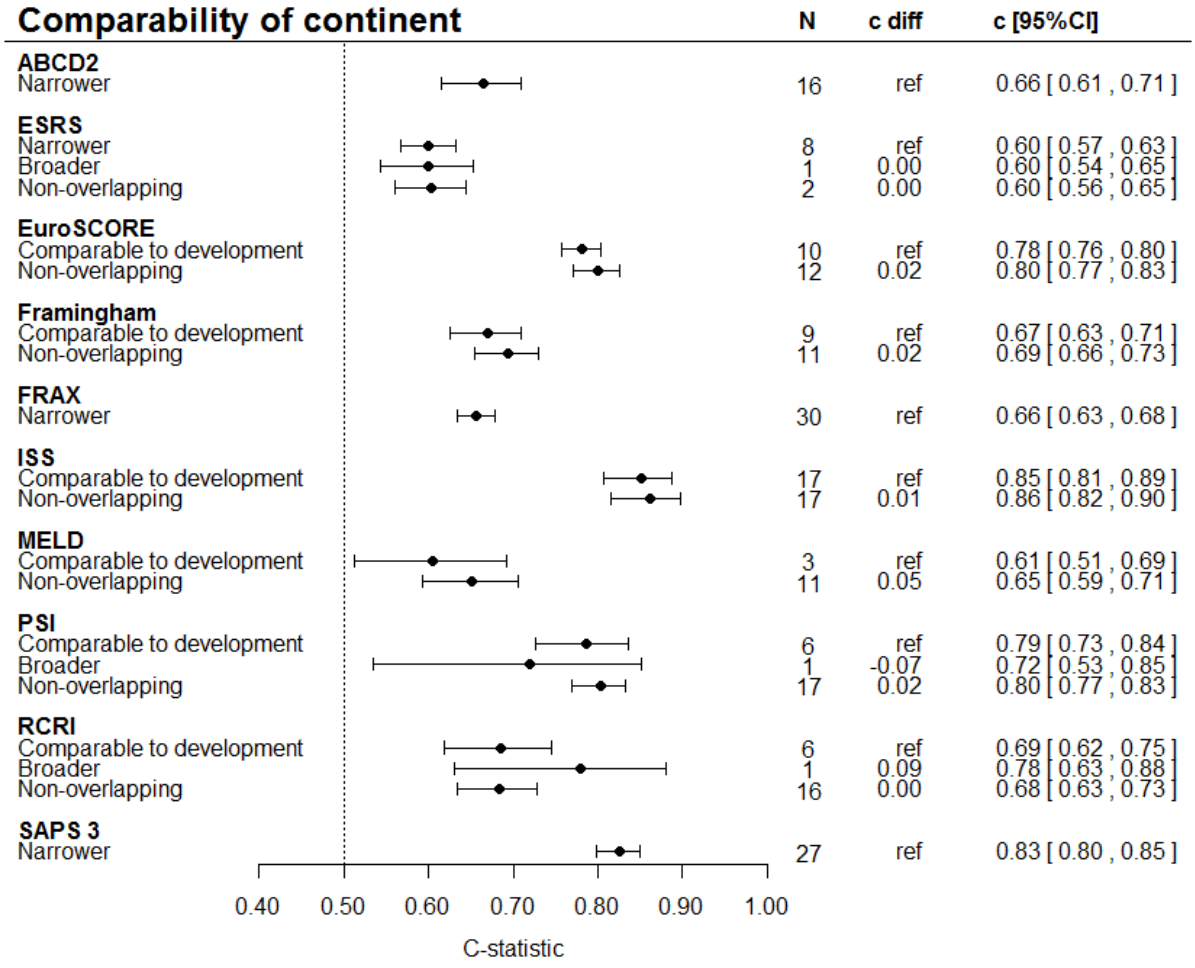


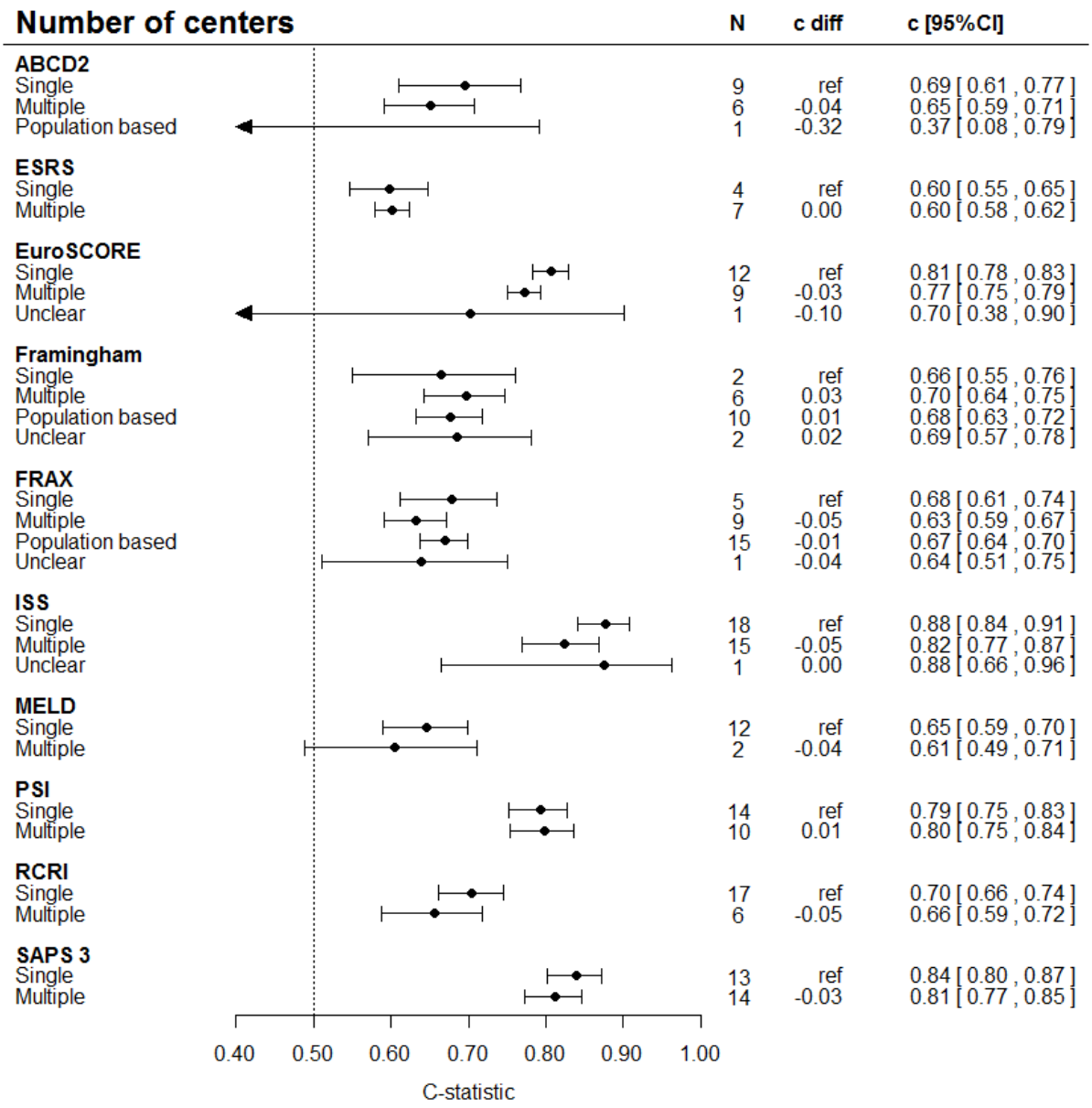

Validation by independent investigators
N
c diff
c [95%CI]
ABCD2
No 6 ref 0.68 [ 0.61 , 0.74 ]
Yes 10 -0.03 0.65 [ 0.58 , 0.72 ]
ESRS
No 3 ref 0.61 [ 0.58 , 0.64 ]
Yes 8 -0.02 0.60 [ 0.58 , 0.62 ]
EuroSCORE
No 3 ref 0.76 [ 0.68 , 0.83 ]
Yes 19 0.03 0.79 [ 0.77 , 0.81 ]
Framingham
No 8 ref 0.66 [ 0.62 , 0.71 ]
Yes 12 0.03 0.69 [ 0.66 , 0.73 ]
FRAX
No 17 ref 0.65 [ 0.62 , 0.67 ]
Yes 13 0.03 0.67 [ 0.64 , 0.71 ]
ISS
No 2 ref 0.88 [ 0.75 , 0.95 ]
Yes 32 -0.03 0.85 [ 0.82 , 0.88 ]
MELD
Yes 14 ref 0.64 [ 0.59 , 0.68 ]
PSI
No 7 ref 0.82 [ 0.77 , 0.86 ]
Yes 17 -0.04 0.78 [ 0.75 , 0.81 ]
RCRI
No 1 ref 0.81 [ 0.66 , 0.90 ]
Yes 22 -0.12 0.68 [ 0.65 , 0.72 ]
SAPS 3
No 2 ref 0.83 [ 0.71 , 0.90 ]
Yes 25 0.00 0.83 [ 0.80 , 0.85 ]
0.40
0.50
0.60
0.70
0.80
0.90
1.00
C-statistic

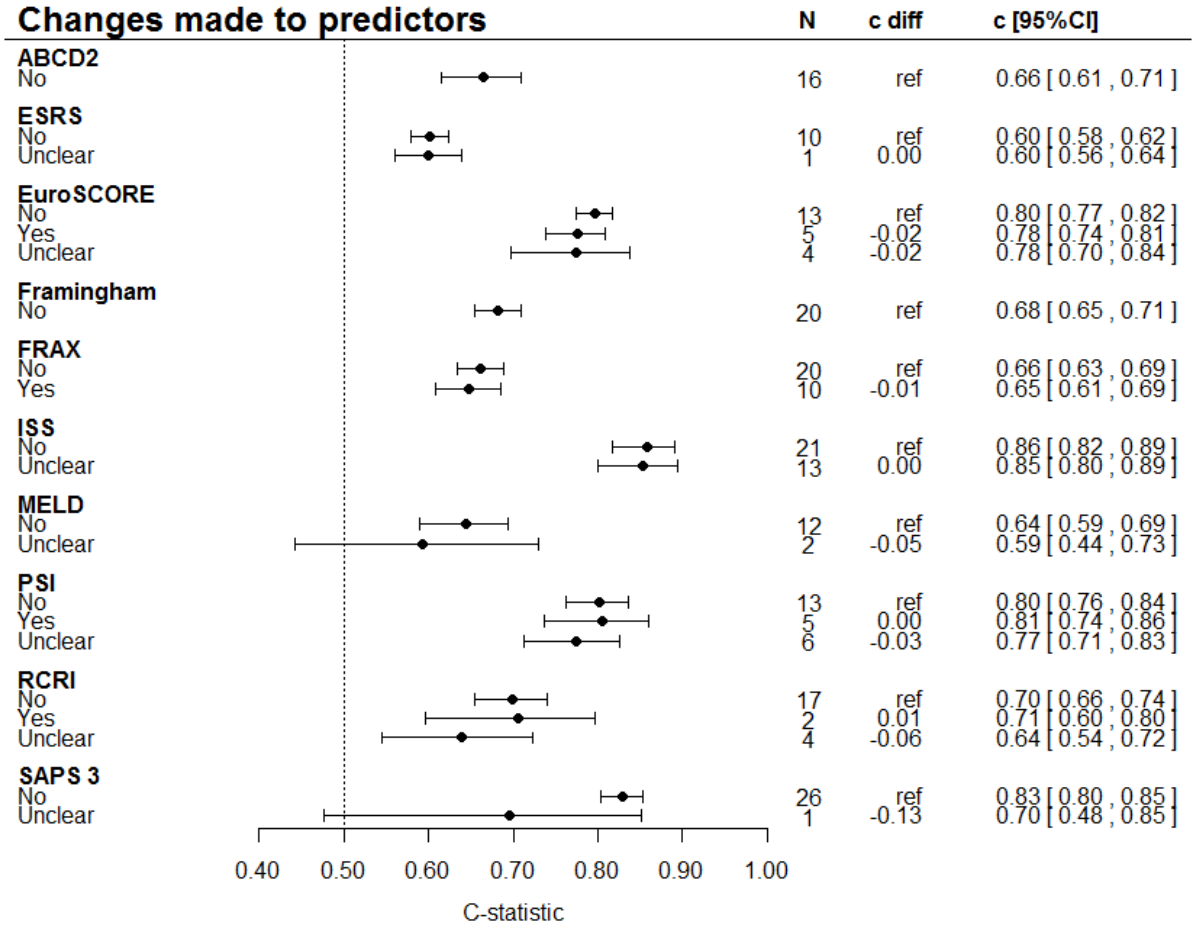

Comparability of eligibility criteria
N
c diff
c [95%CI]
ABCD2
Narrower 6 ref 0.64 [ 0.57 , 0.71 ]
Comparable 4 0.04 0.68 [ 0.59 , 0.76 ]
Mixture 5 -0.01 0.64 [ 0.50 , 0.75 ]
Broader 1 0.15 0.79 [ 0.60 , 0.90 ]
ESRS
Narrower 2 ref 0.59 [ 0.54 , 0.64 ]
Mixture 9 0.01 0.60 [ 0.58 , 0.62 ]
EuroSCORE
Narrower 18 ref 0.78 [ 0.77 , 0.80 ]
Comparable 2 -0.01 0.77 [ 0.70 , 0.83 ]
Broader 2 0.04 0.83 [ 0.78 , 0.87 ]
Framingham
Narrower 15 ref 0.66 [ 0.64 , 0.69 ]
Comparable 3 0.04 0.70 [ 0.66 , 0.75 ]
Mixture 2 0.12 0.79 [ 0.73 , 0.84 ]
FRAX
Narrower 28 ref 0.65 [ 0.63 , 0.67 ]
Comparable 2 0.07 0.72 [ 0.64 , 0.79 ]
ISS
Narrower 22 ref 0.84 [ 0.80 , 0.87 ]
Comparable 5 0.05 0.89 [ 0.83 , 0.94 ]
Mixture 3 -0.06 0.78 [ 0.60 , 0.89 ]
Broader 2 0.10 0.94 [ 0.86 , 0.97 ]
Unclear 2 0.04 0.88 [ 0.76 , 0.95 ]
MELD
Non-overlapping 14 ref 0.64 [ 0.59 , 0.68 ]
PSI
Narrower 4 ref 0.78 [ 0.70 , 0.84 ]
Comparable 1 0.05 0.83 [ 0.70 , 0.91 ]
Mixture 16 0.01 0.79 [ 0.75 , 0.82 ]
Broader 3 0.07 0.85 [ 0.77 , 0.90 ]
RCRI
Narrower 4 ref 0.67 [ 0.58 , 0.75 ]
Comparable 3 0.10 0.77 [ 0.70 , 0.84 ]
Mixture 11 0.00 0.67 [ 0.62 , 0.73 ]
Broader 5 0.00 0.67 [ 0.59 , 0.73 ]
SAPS 3
Narrower 20 ref 0.82 [ 0.78 , 0.85 ]
Comparable 7 0.03 0.85 [ 0.80 , 0.89 ]
0.40
0.50
0.60
0.70
0.80
0.90
1.00
C-statistic

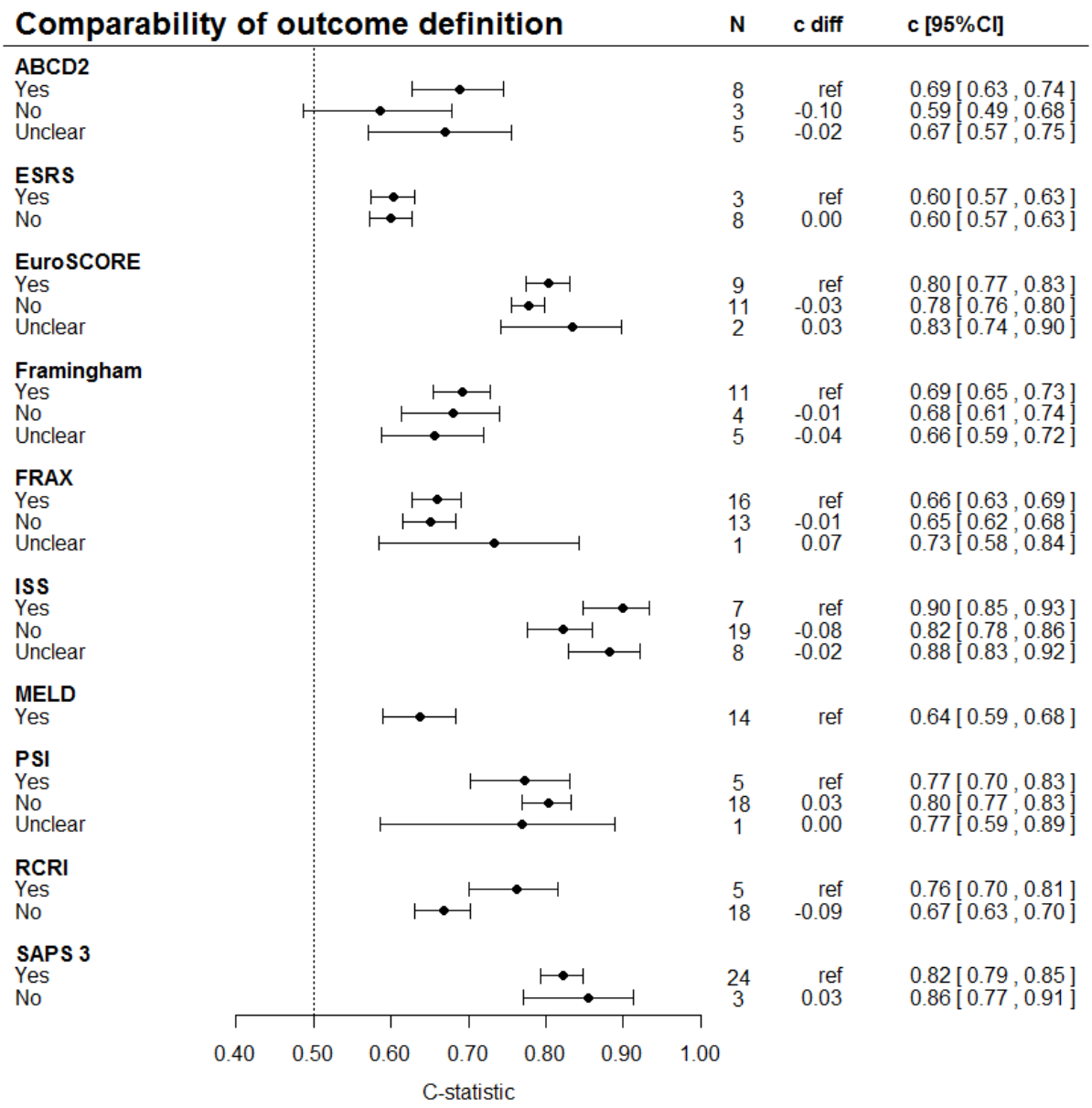

Comparability of age eligibility criteria
N
c diff
c [95%CI]
ABCD2*
Narrower 1 ref 0.62 [ 0.35 , 0.83 ]
Comparable 15 0.05 0.67 [ 0.62 , 0.72 ]
ESRS*
Narrower 2 ref 0.60 [ 0.57 , 0.64 ]
Broader 9 0.00 0.60 [ 0.58 , 0.62 ]
EuroSCORE*
Comparable 22 ref 0.79 [ 0.77 , 0.81 ]
Framingham*
Narrower 14 ref 0.68 [ 0.64 , 0.71 ]
Mixture 6 0.02 0.70 [ 0.65 , 0.75 ]
FRAX*
Narrower 9 ref 0.68 [ 0.63 , 0.72 ]
Comparable 1 0.06 0.73 [ 0.58 , 0.84 ]
Mixture 16 -0.03 0.64 [ 0.61 , 0.67 ]
Broader 4 -0.01 0.66 [ 0.60 , 0.72 ]
ISS
Narrower 13 ref 0.82 [ 0.77 , 0.87 ]
Comparable 21 0.05 0.87 [ 0.84 , 0.90 ]
MELD
Narrower 10 ref 0.61 [ 0.56 , 0.66 ]
Comparable 4 0.10 0.70 [ 0.63 , 0.76 ]
PSI*
Narrower 2 ref 0.76 [ 0.64 , 0.84 ]
Comparable 15 0.06 0.81 [ 0.78 , 0.84 ]
Broader 7 0.01 0.77 [ 0.71 , 0.82 ]
RCRI
Narrower 1 ref 0.80 [ 0.65 , 0.90 ]
Comparable 4 -0.03 0.77 [ 0.72 , 0.82 ]
Broader 18 -0.14 0.66 [ 0.63 , 0.69 ]
SAPS 3*
Narrower 17 ref 0.83 [ 0.79 , 0.86 ]
Comparable 4 -0.01 0.81 [ 0.73 , 0.88 ]
Broader 6 0.00 0.83 [ 0.77 , 0.88 ]
0.40
0.50
0.60
0.70
0.80
0.90
1.00
C-statistic

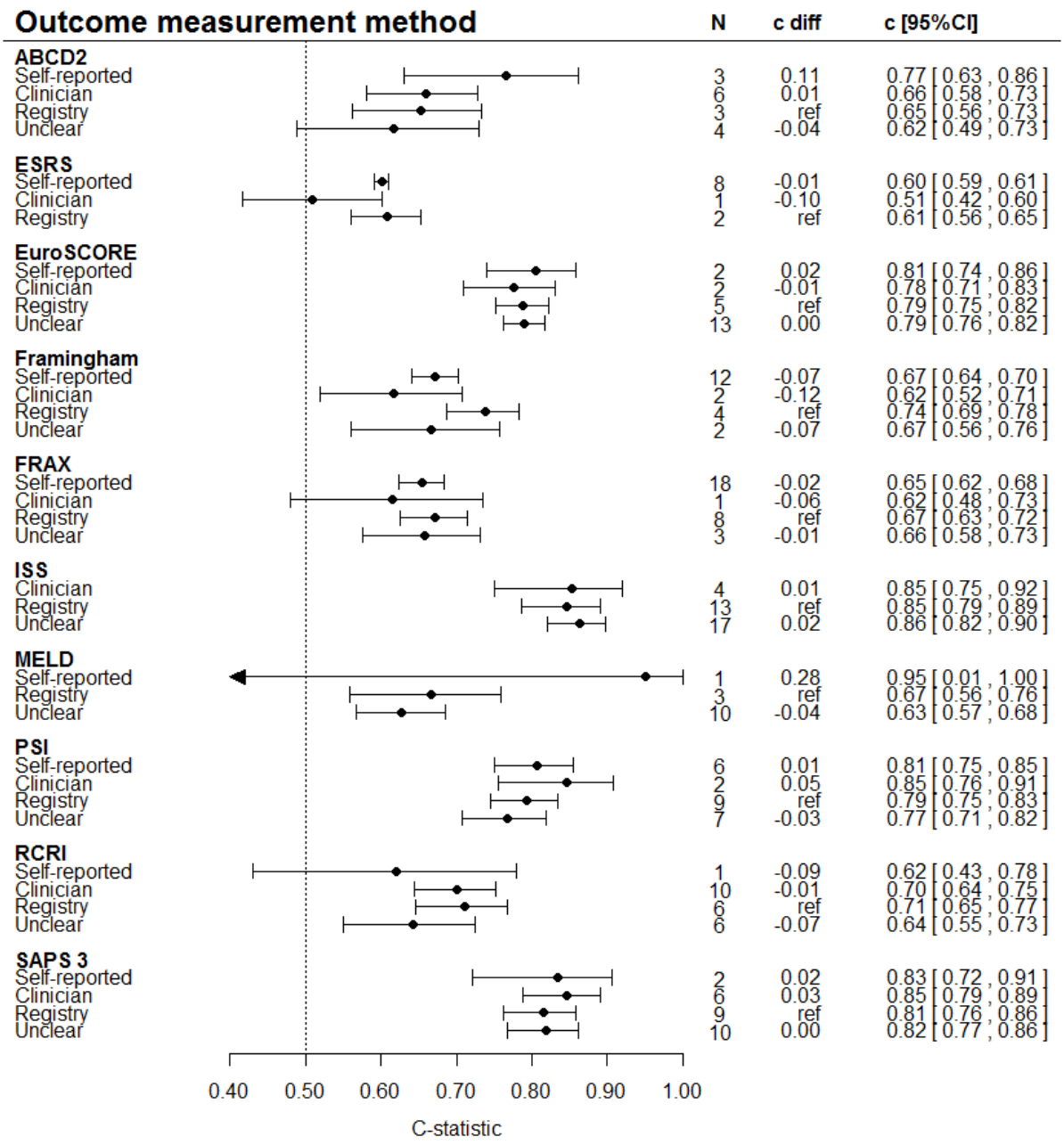

Comparability of setting
N
c diff
c [95%CI]
ABCD2
Narrower 15 0.30 0.67 [ 0.62 , 0.71 ]
Comparable 1 ref 0.37 [ 0.08 , 0.79 ]
ESRS
Non-overlapping 11 ref 0.60 [ 0.58 , 0.62 ]
EuroSCORE
Comparable 16 ref 0.79 [ 0.76 , 0.81 ]
Broader 2 -0.02 0.77 [ 0.00 , 1.00 ]
Non-overlapping 3 0.02 0.80 [ 0.76 , 0.84 ]
Unclear 1 -0.09 0.70 [ 0.38 , 0.90 ]
Framingham
Comparable 13 ref 0.68 [ 0.64 , 0.71 ]
Non-overlapping 7 0.01 0.69 [ 0.64 , 0.73 ]
FRAX
Comparable 17 ref 0.66 [ 0.63 , 0.69 ]
Non-overlapping 10 -0.01 0.66 [ 0.61 , 0.70 ]
Unclear 3 -0.03 0.63 [ 0.56 , 0.70 ]
ISS
Comparable 18 ref 0.86 [ 0.82 , 0.90 ]
Broader 4 -0.06 0.80 [ 0.66 , 0.90 ]
Non-overlapping 12 -0.01 0.86 [ 0.80 , 0.90 ]
MELD
Comparable 6 ref 0.64 [ 0.56 , 0.72 ]
Broader 1 -0.09 0.55 [ 0.38 , 0.71 ]
Non-overlapping 6 0.01 0.66 [ 0.57 , 0.73 ]
Unclear 1 -0.01 0.63 [ 0.46 , 0.77 ]
PSI
Comparable 18 ref 0.80 [ 0.77 , 0.83 ]
Broader 2 0.00 0.80 [ 0.71 , 0.87 ]
Non-overlapping 4 -0.06 0.75 [ 0.67 , 0.81 ]
RCRI
Comparable 9 ref 0.73 [ 0.68 , 0.77 ]
Broader 2 -0.12 0.61 [ 0.49 , 0.71 ]
Non-overlapping 12 -0.06 0.67 [ 0.62 , 0.72 ]
SAPS 3
Comparable 4 ref 0.84 [ 0.76 , 0.89 ]
Broader 2 -0.04 0.80 [ 0.66 , 0.89 ]
Non-overlapping 17 -0.01 0.82 [ 0.79 , 0.86 ]
Unclear 4 0.00 0.83 [ 0.76 , 0.89 ]
0.40
0.50
0.60
0.70
0.80
0.90
1.00
C-statistic

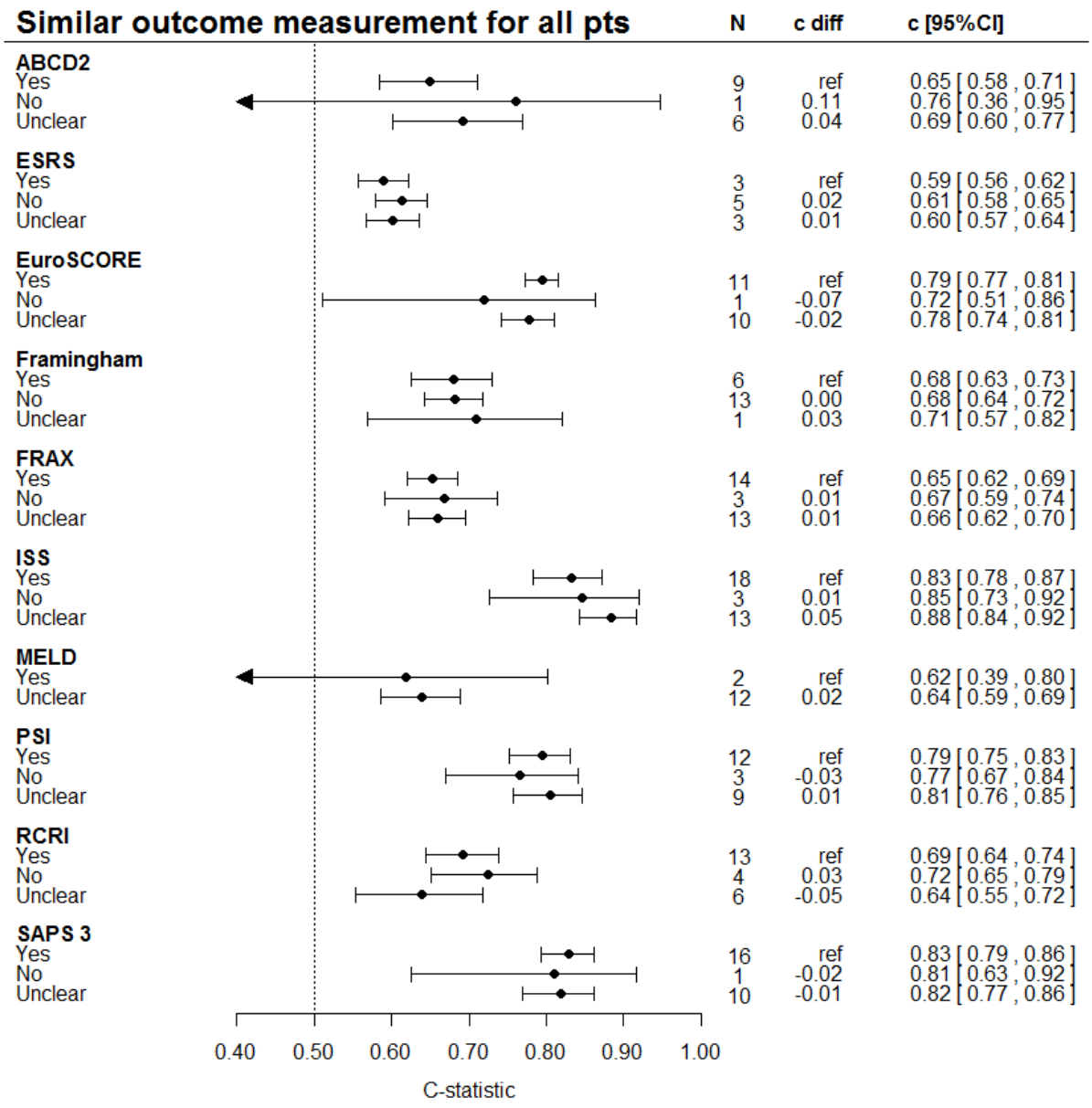

Comparability of continent
N
c diff
c [95%CI]
ABCD2
Narrower
16
ref
0.66 [ 0.61 , 0.71 ]
ESRS
Narrower
8
ref
0.60 [ 0.57 , 0.63 ]
Broader
1
0.00
0.60 [ 0.54 , 0.65 ]
Non-overlapping
2
0.00
0.60 [ 0.56 , 0.65 ]
EuroSCORE
Comparable to development
10
ref
0.78 [ 0.76 , 0.80 ]
Non-overlapping
12
0.02
0.80 [ 0.77 , 0.83 ]
Framingham
Comparable to development
9
ref
0.67 [ 0.63 , 0.71 ]
Non-overlapping
11
0.02
0.69 [ 0.66 , 0.73 ]
FRAX
Narrower
30
ref
0.66 [ 0.63 , 0.68 ]
ISS
Comparable to development
17
ref
0.85 [ 0.81 , 0.89 ]
Non-overlapping
17
0.01
0.86 [ 0.82 , 0.90 ]
MELD
Comparable to development
3
ref
0.61 [ 0.51 , 0.69 ]
Non-overlapping
11
0.05
0.65 [ 0.59 , 0.71 ]
PSI
Comparable to development
6
ref
0.79 [ 0.73 , 0.84 ]
Broader
1
-0.07
0.72 [ 0.53 , 0.85 ]
Non-overlapping
17
0.02
0.80 [ 0.77 , 0.83 ]
RCRI
Comparable to development
6
ref
0.69 [ 0.62 , 0.75 ]
Broader
1
0.09
0.78 [ 0.63 , 0.88 ]
Non-overlapping
16
0.00
0.68 [ 0.63 , 0.73 ]
SAPS 3
Narrower
27
ref
0.83 [ 0.80 , 0.85 ]
0.40
0.50
0.60
0.70
0.80
0.90
1.00
C-statistic

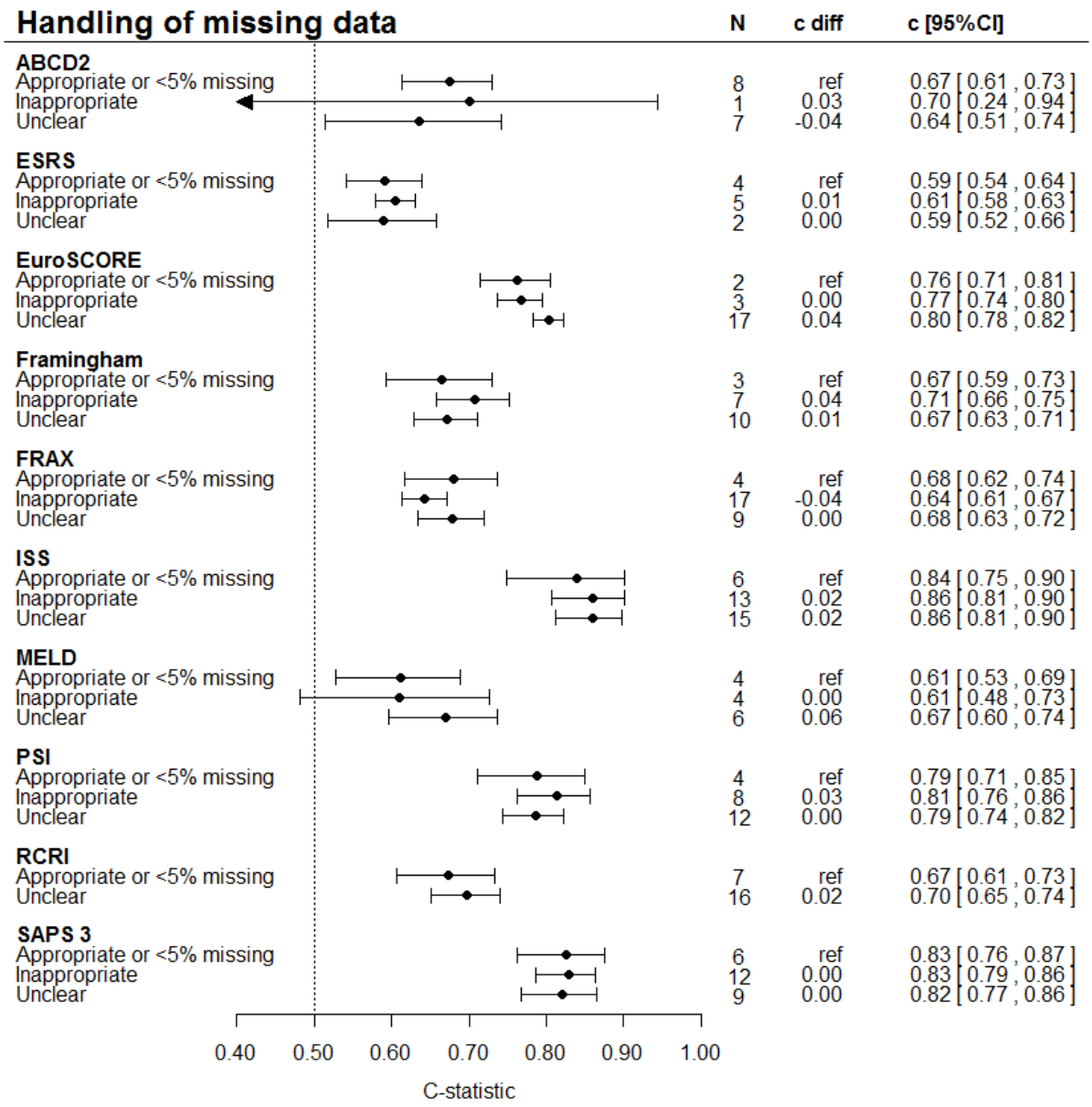

Number of centers
N
c diff
c [95%CI]
ABCD2
Single 9 ref 0.69 [ 0.61 , 0.77 ]
Multiple 6 -0.04 0.65 [ 0.59 , 0.71 ]
Population based 1 -0.32 0.37 [ 0.08 , 0.79 ]
ESRS
Single 4 ref 0.60 [ 0.55 , 0.65 ]
Multiple 7 0.00 0.60 [ 0.58 , 0.62 ]
EuroSCORE
Single 12 ref 0.81 [ 0.78 , 0.83 ]
Multiple 9 -0.03 0.77 [ 0.75 , 0.79 ]
Unclear 1 -0.10 0.70 [ 0.38 , 0.90 ]
Framingham
Single 2 ref 0.66 [ 0.55 , 0.76 ]
Multiple 6 0.03 0.70 [ 0.64 , 0.75 ]
Population based 10 0.01 0.68 [ 0.63 , 0.72 ]
Unclear 2 0.02 0.69 [ 0.57 , 0.78 ]
FRAX
Single 5 ref 0.68 [ 0.61 , 0.74 ]
Multiple 9 -0.05 0.63 [ 0.59 , 0.67 ]
Population based 15 -0.01 0.67 [ 0.64 , 0.70 ]
Unclear 1 -0.04 0.64 [ 0.51 , 0.75 ]
ISS
Single 18 ref 0.88 [ 0.84 , 0.91 ]
Multiple 15 -0.05 0.82 [ 0.77 , 0.87 ]
Unclear 1 0.00 0.88 [ 0.66 , 0.96 ]
MELD
Single 12 ref 0.65 [ 0.59 , 0.70 ]
Multiple 2 -0.04 0.61 [ 0.49 , 0.71 ]
PSI
Single 14 ref 0.79 [ 0.75 , 0.83 ]
Multiple 10 0.01 0.80 [ 0.75 , 0.84 ]
RCRI
Single 17 ref 0.70 [ 0.66 , 0.74 ]
Multiple 6 -0.05 0.66 [ 0.59 , 0.72 ]
SAPS 3
Single 13 ref 0.84 [ 0.80 , 0.87 ]
Multiple 14 -0.03 0.81 [ 0.77 , 0.85 ]
0.40
0.50
0.60
0.70
0.80
0.90
1.00
C-statistic


Changes made to predictors
N
c diff
c [95%CI]
ABCD2
No
16
ref
0.66 [ 0.61 , 0.71 ]
ESRS
No
10
ref
0.60 [ 0.58 , 0.62 ]
Unclear
1
0.00
0.60 [ 0.56 , 0.64 ]
EuroSCORE
No
13
ref
0.80 [ 0.77 , 0.82 ]
Yes
5
-0.02
0.78 [ 0.74 , 0.81 ]
Unclear
4
-0.02
0.78 [ 0.70 , 0.84 ]
Framingham
No
20
ref
0.68 [ 0.65 , 0.71 ]
FRAX
No
20
ref
0.66 [ 0.63 , 0.69 ]
Yes
10
-0.01
0.65 [ 0.61 , 0.69 ]
ISS
No
21
ref
0.86 [ 0.82 , 0.89 ]
Unclear
13
0.00
0.85 [ 0.80 , 0.89 ]
MELD
No
12
ref
0.64 [ 0.59 , 0.69 ]
Unclear
2
-0.05
0.59 [ 0.44 , 0.73 ]
PSI
No
13
ref
0.80 [ 0.76 , 0.84 ]
Yes
5
0.00
0.81 [ 0.74 , 0.86 ]
Unclear
6
-0.03
0.77 [ 0.71 , 0.83 ]
RCRI
No
17
ref
0.70 [ 0.66 , 0.74 ]
Yes
2
0.01
0.71 [ 0.60 , 0.80 ]
Unclear
4
-0.06
0.64 [ 0.54 , 0.72 ]
SAPS 3
No
26
ref
0.83 [ 0.80 , 0.85 ]
Unclear
1
-0.13
0.70 [ 0.48 , 0.85 ]
0.40
0.50
0.60
0.70
0.80
0.90
1.00
C-statistic


Comparability of outcome definition
N
c diff
c [95%CI]
ABCD2
Yes 8 ref 0.69 [ 0.63 , 0.74 ]
No 3 -0.10 0.59 [ 0.49 , 0.68 ]
Unclear 5 -0.02 0.67 [ 0.57 , 0.75 ]
ESRS
Yes 3 ref 0.60 [ 0.57 , 0.63 ]
No 8 0.00 0.60 [ 0.57 , 0.63 ]
EuroSCORE
Yes 9 ref 0.80 [ 0.77 , 0.83 ]
No 11 -0.03 0.78 [ 0.76 , 0.80 ]
Unclear 2 0.03 0.83 [ 0.74 , 0.90 ]
Framingham
Yes 11 ref 0.69 [ 0.65 , 0.73 ]
No 4 -0.01 0.68 [ 0.61 , 0.74 ]
Unclear 5 -0.04 0.66 [ 0.59 , 0.72 ]
FRAX
Yes 16 ref 0.66 [ 0.63 , 0.69 ]
No 13 -0.01 0.65 [ 0.62 , 0.68 ]
Unclear 1 0.07 0.73 [ 0.58 , 0.84 ]
ISS
Yes 7 ref 0.90 [ 0.85 , 0.93 ]
No 19 -0.08 0.82 [ 0.78 , 0.86 ]
Unclear 8 -0.02 0.88 [ 0.83 , 0.92 ]
MELD
Yes 14 ref 0.64 [ 0.59 , 0.68 ]
PSI
Yes 5 ref 0.77 [ 0.70 , 0.83 ]
No 18 0.03 0.80 [ 0.77 , 0.83 ]
Unclear 1 0.00 0.77 [ 0.59 , 0.89 ]
RCRI
Yes 5 ref 0.76 [ 0.70 , 0.81 ]
No 18 -0.09 0.67 [ 0.63 , 0.70 ]
SAPS 3
Yes 24 ref 0.82 [ 0.79 , 0.85 ]
No 3 0.03 0.86 [ 0.77 , 0.91 ]
0.40
0.50
0.60
0.70
0.80
0.90
1.00
C-statistic


Outcome measurement method
N
c diff
c [95%CI]
ABCD2
Self-reported 3 0.11 0.77 [ 0.63 , 0.86 ]
Clinician 6 0.01 0.66 [ 0.58 , 0.73 ]
Registry 3 ref 0.65 [ 0.56 , 0.73 ]
Unclear 4 -0.04 0.62 [ 0.49 , 0.73 ]
ESRS
Self-reported 8 -0.01 0.60 [ 0.59 , 0.61 ]
Clinician 1 -0.10 0.51 [ 0.42 , 0.60 ]
Registry 2 ref 0.61 [ 0.56 , 0.65 ]
EuroSCORE
Self-reported 2 0.02 0.81 [ 0.74 , 0.86 ]
Clinician 2 -0.01 0.78 [ 0.71 , 0.83 ]
Registry 5 ref 0.79 [ 0.75 , 0.82 ]
Unclear 13 0.00 0.79 [ 0.76 , 0.82 ]
Framingham
Self-reported 12 -0.07 0.67 [ 0.64 , 0.70 ]
Clinician 2 -0.12 0.62 [ 0.52 , 0.71 ]
Registry 4 ref 0.74 [ 0.69 , 0.78 ]
Unclear 2 -0.07 0.67 [ 0.56 , 0.76 ]
FRAX
Self-reported 18 -0.02 0.65 [ 0.62 , 0.68 ]
Clinician 1 -0.06 0.62 [ 0.48 , 0.73 ]
Registry 8 ref 0.67 [ 0.63 , 0.72 ]
Unclear 3 -0.01 0.66 [ 0.58 , 0.73 ]
ISS
Clinician 4 0.01 0.85 [ 0.75 , 0.92 ]
Registry 13 ref 0.85 [ 0.79 , 0.89 ]
Unclear 17 0.02 0.86 [ 0.82 , 0.90 ]
MELD
Self-reported 1 0.28 0.95 [ 0.01 , 1.00 ]
Registry 3 ref 0.67 [ 0.56 , 0.76 ]
Unclear 10 -0.04 0.63 [ 0.57 , 0.68 ]
PSI
Self-reported 6 0.01 0.81 [ 0.75 , 0.85 ]
Clinician 2 0.05 0.85 [ 0.76 , 0.91 ]
Registry 9 ref 0.79 [ 0.75 , 0.83 ]
Unclear 7 -0.03 0.77 [ 0.71 , 0.82 ]
RCRI
Self-reported 1 -0.09 0.62 [ 0.43 , 0.78 ]
Clinician 10 -0.01 0.70 [ 0.64 , 0.75 ]
Registry 6 ref 0.71 [ 0.65 , 0.77 ]
Unclear 6 -0.07 0.64 [ 0.55 , 0.73 ]
SAPS 3
Self-reported 2 0.02 0.83 [ 0.72 , 0.91 ]
Clinician 6 0.03 0.85 [ 0.79 , 0.89 ]
Registry 9 ref 0.81 [ 0.76 , 0.86 ]
Unclear 10 0.00 0.82 [ 0.77 , 0.86 ]
0.40
0.50
0.60
0.70
0.80
0.90
1.00
C-statistic


Similar outcome measurement for all pts
N
c diff
c [95%CI]
ABCD2
Yes 9 ref 0.65 [ 0.58 , 0.71 ]
No 1 0.11 0.76 [ 0.36 , 0.95 ]
Unclear 6 0.04 0.69 [ 0.60 , 0.77 ]
ESRS
Yes 3 ref 0.59 [ 0.56 , 0.62 ]
No 5 0.02 0.61 [ 0.58 , 0.65 ]
Unclear 3 0.01 0.60 [ 0.57 , 0.64 ]
EuroSCORE
Yes 11 ref 0.79 [ 0.77 , 0.81 ]
No 1 -0.07 0.72 [ 0.51 , 0.86 ]
Unclear 10 -0.02 0.78 [ 0.74 , 0.81 ]
Framingham
Yes 6 ref 0.68 [ 0.63 , 0.73 ]
No 13 0.00 0.68 [ 0.64 , 0.72 ]
Unclear 1 0.03 0.71 [ 0.57 , 0.82 ]
FRAX
Yes 14 ref 0.65 [ 0.62 , 0.69 ]
No 3 0.01 0.67 [ 0.59 , 0.74 ]
Unclear 13 0.01 0.66 [ 0.62 , 0.70 ]
ISS
Yes 18 ref 0.83 [ 0.78 , 0.87 ]
No 3 0.01 0.85 [ 0.73 , 0.92 ]
Unclear 13 0.05 0.88 [ 0.84 , 0.92 ]
MELD
Yes 2 ref 0.62 [ 0.39 , 0.80 ]
Unclear 12 0.02 0.64 [ 0.59 , 0.69 ]
PSI
Yes 12 ref 0.79 [ 0.75 , 0.83 ]
No 3 -0.03 0.77 [ 0.67 , 0.84 ]
Unclear 9 0.01 0.81 [ 0.76 , 0.85 ]
RCRI
Yes 13 ref 0.69 [ 0.64 , 0.74 ]
No 4 0.03 0.72 [ 0.65 , 0.79 ]
Unclear 6 -0.05 0.64 [ 0.55 , 0.72 ]
SAPS 3
Yes 16 ref 0.83 [ 0.79 , 0.86 ]
No 1 -0.02 0.81 [ 0.63 , 0.92 ]
Unclear 10 -0.01 0.82 [ 0.77 , 0.86 ]
0.40 0.50 0.60 0.70 0.80 0.90 1.00
C-statistic




C-statistic for categories of study characteristics, pooled using univariable meta-regression analyses per systematic review. N represents the number of external validation studies in a specific category. C diff represents the difference in c-statistic with regard to a reference category (indicated with 'ref'). Dev: development, val: validation, incr: incremental value, pts: patients.
*Models contain age as predictor

Figure S4: Associations between categorical variables and OE ratio

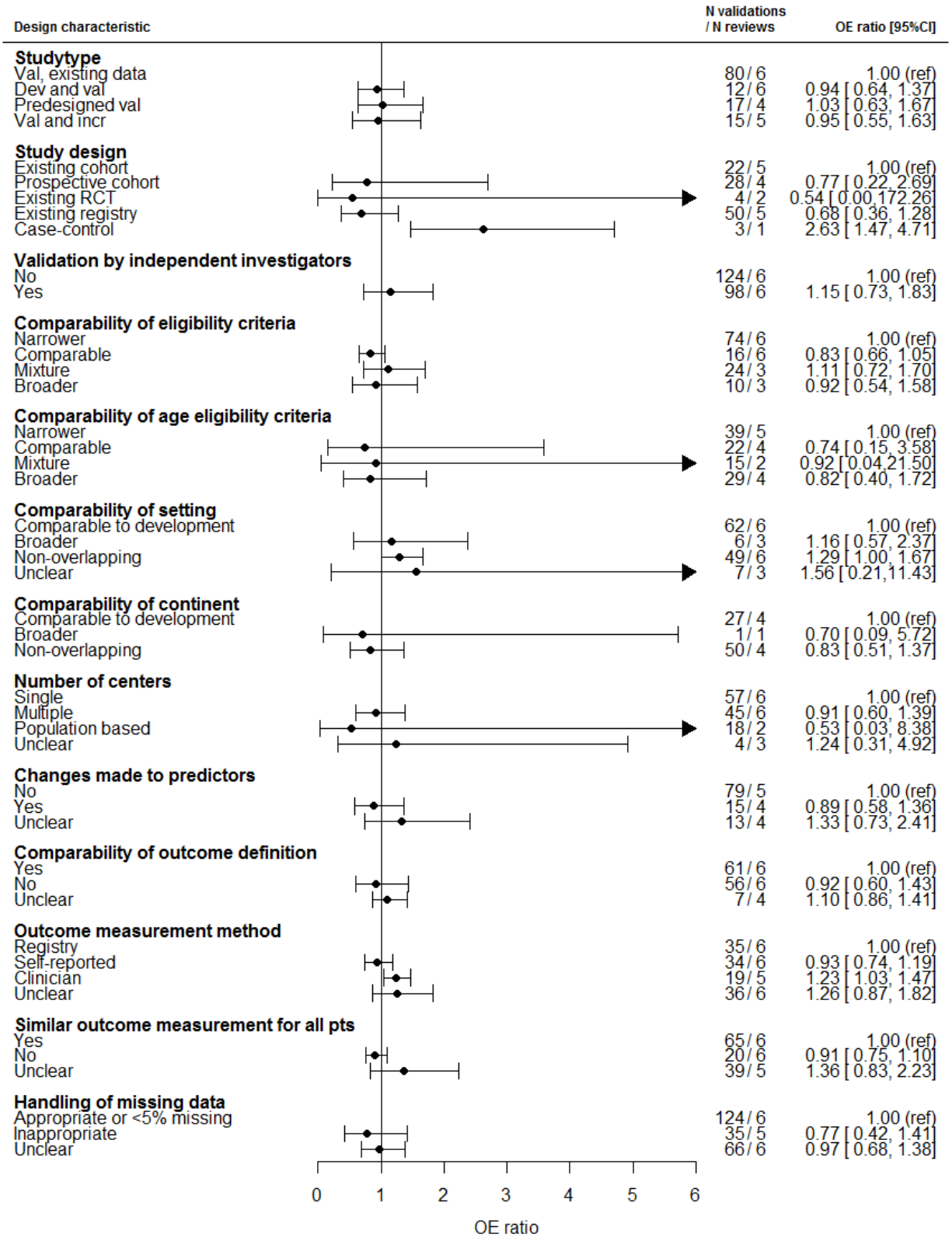
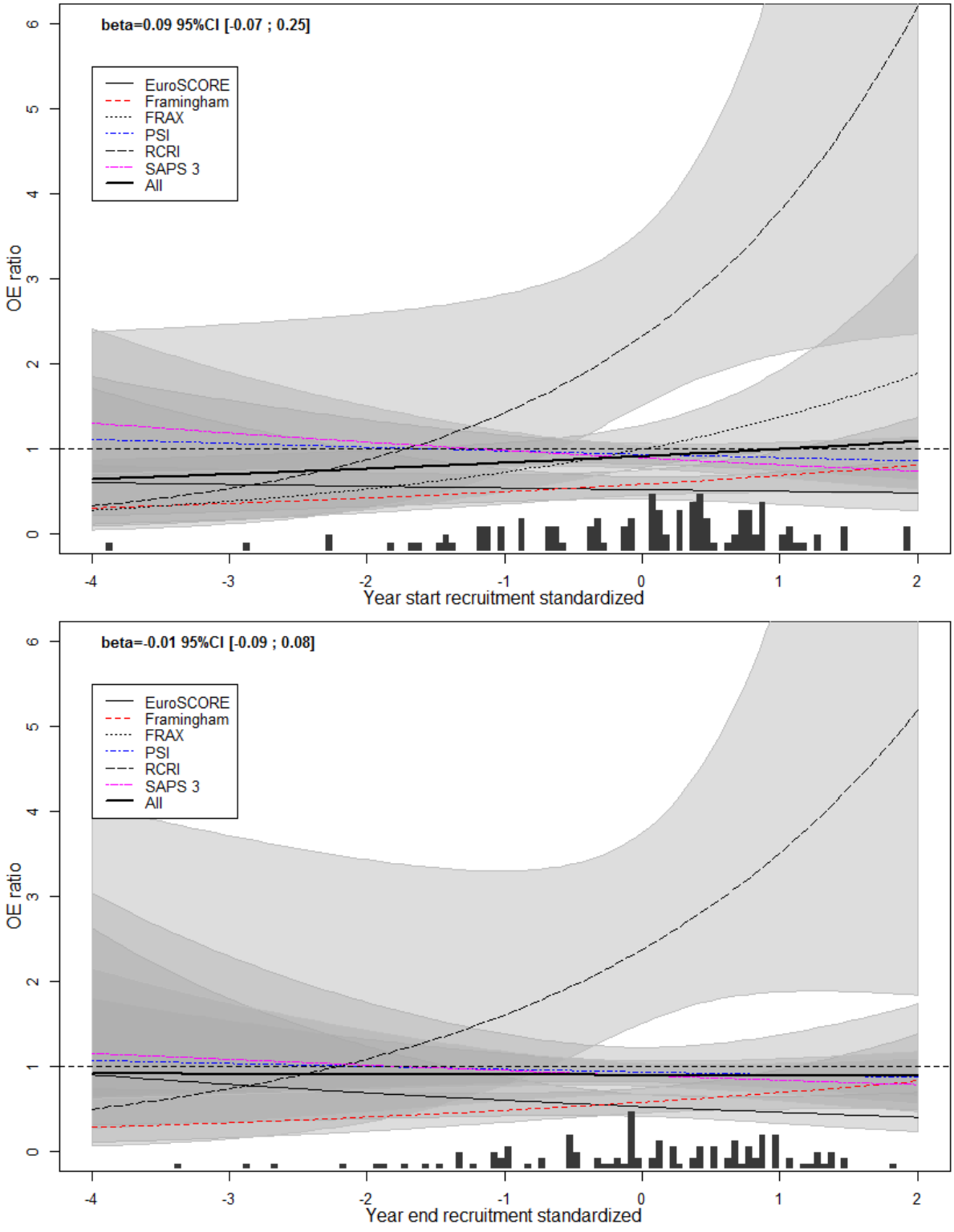
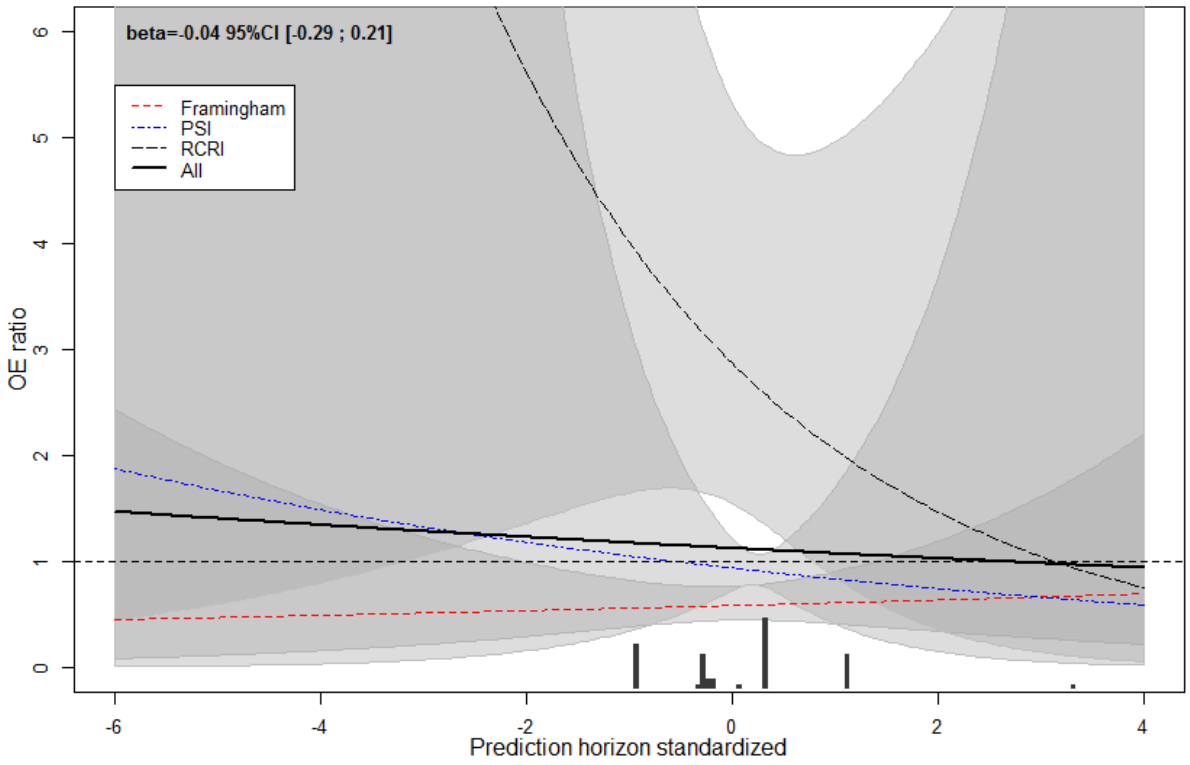
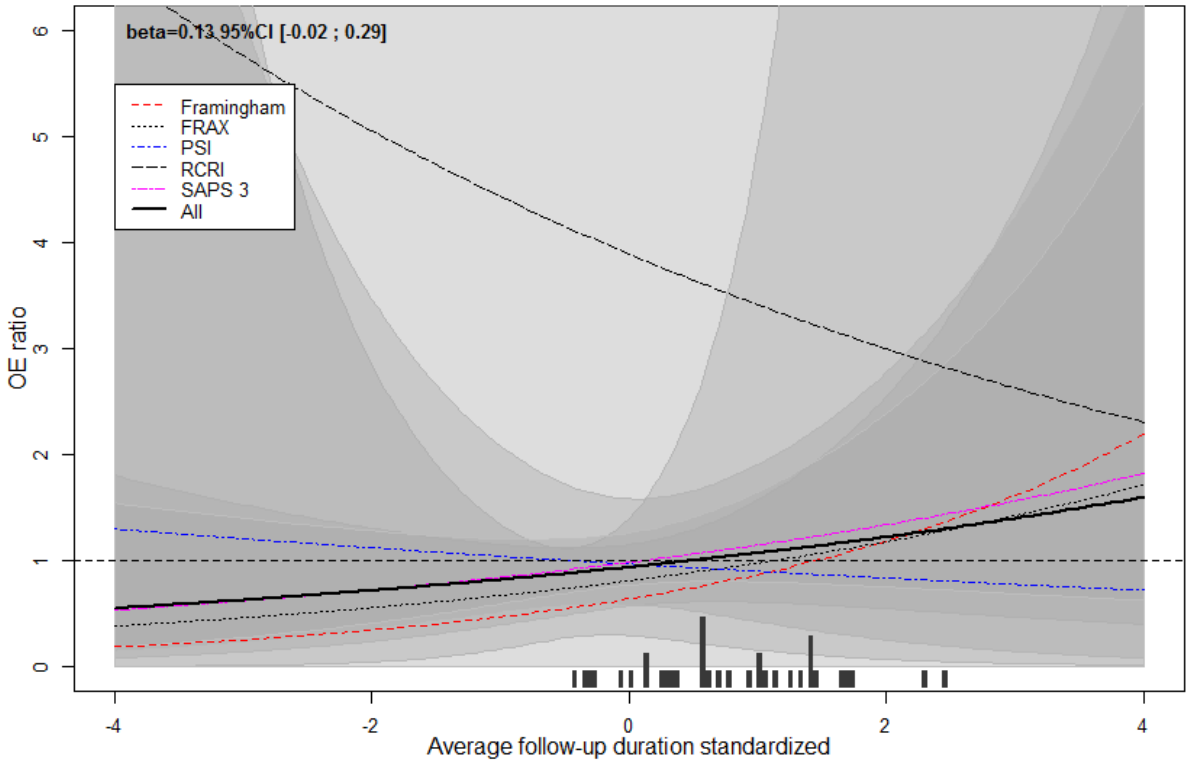
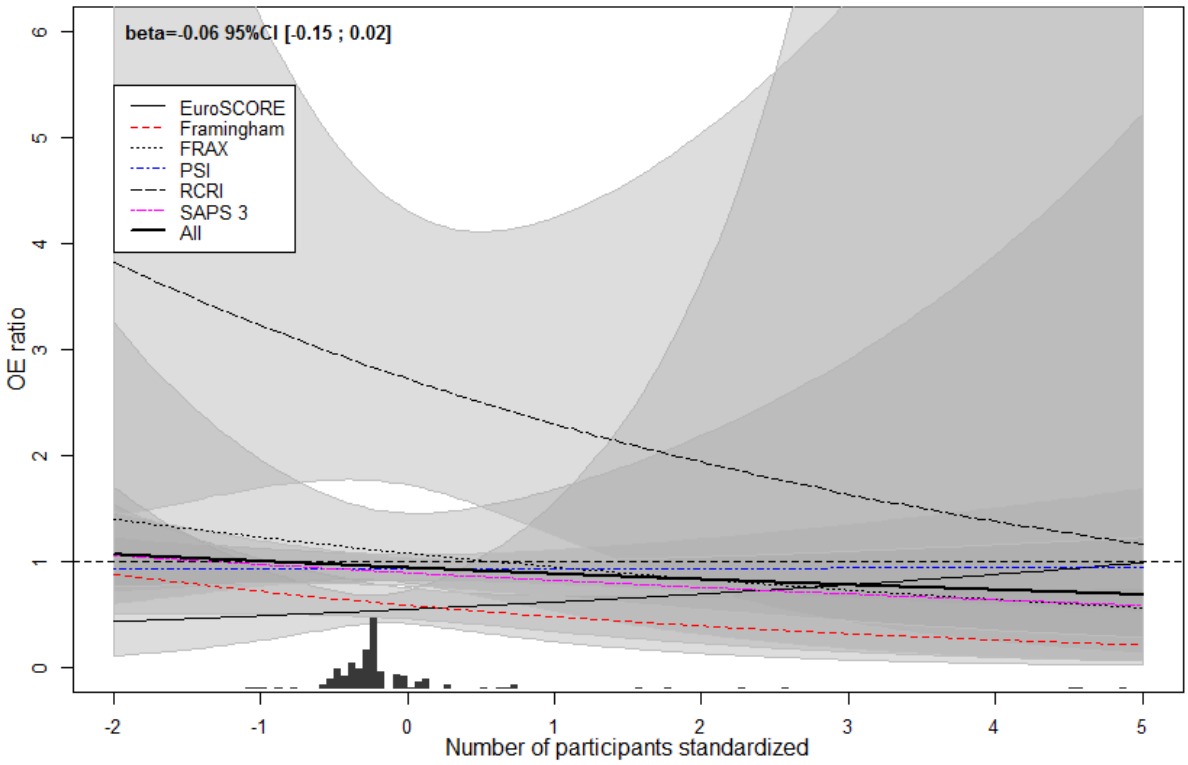


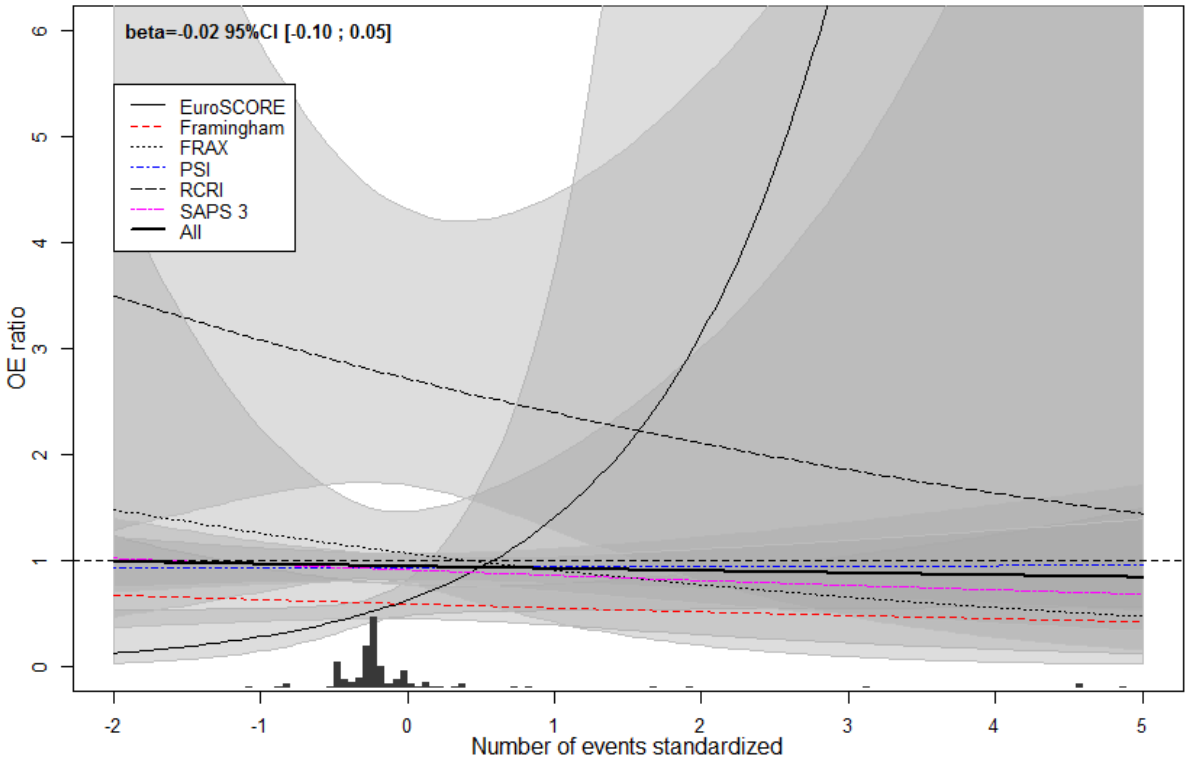
Figure S5: Associations between continuous variables and OE ratio

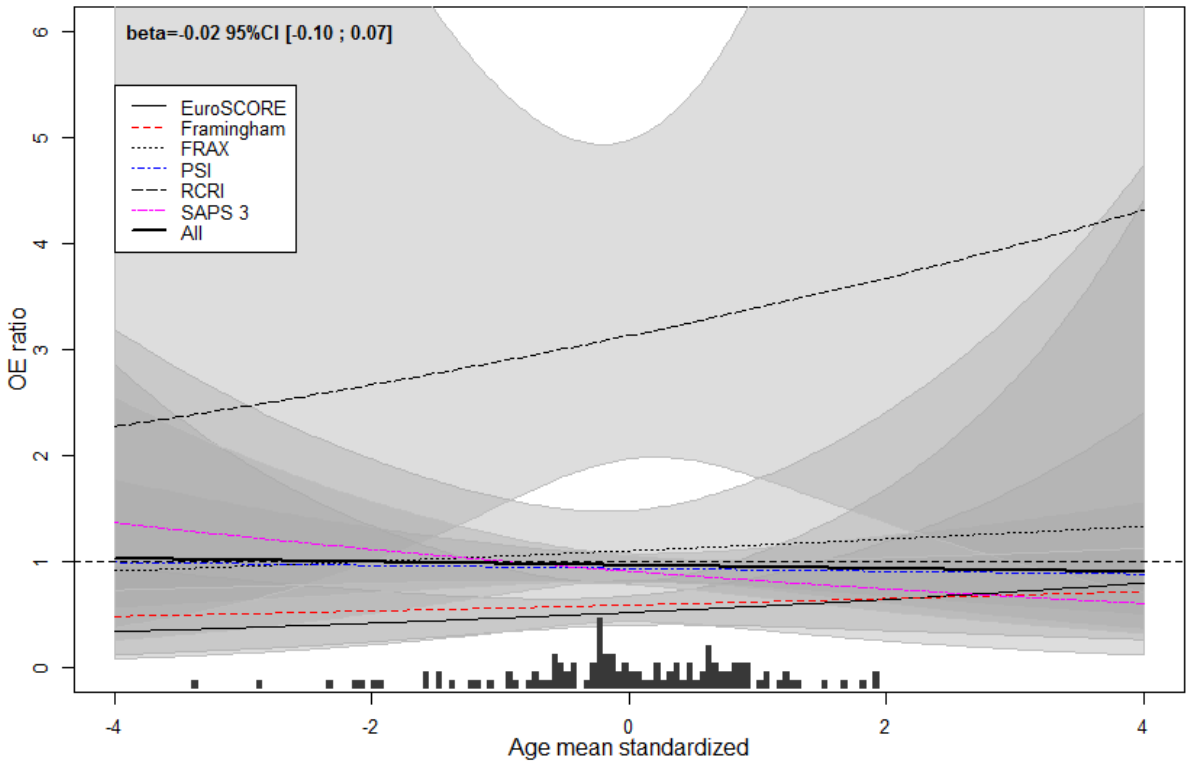


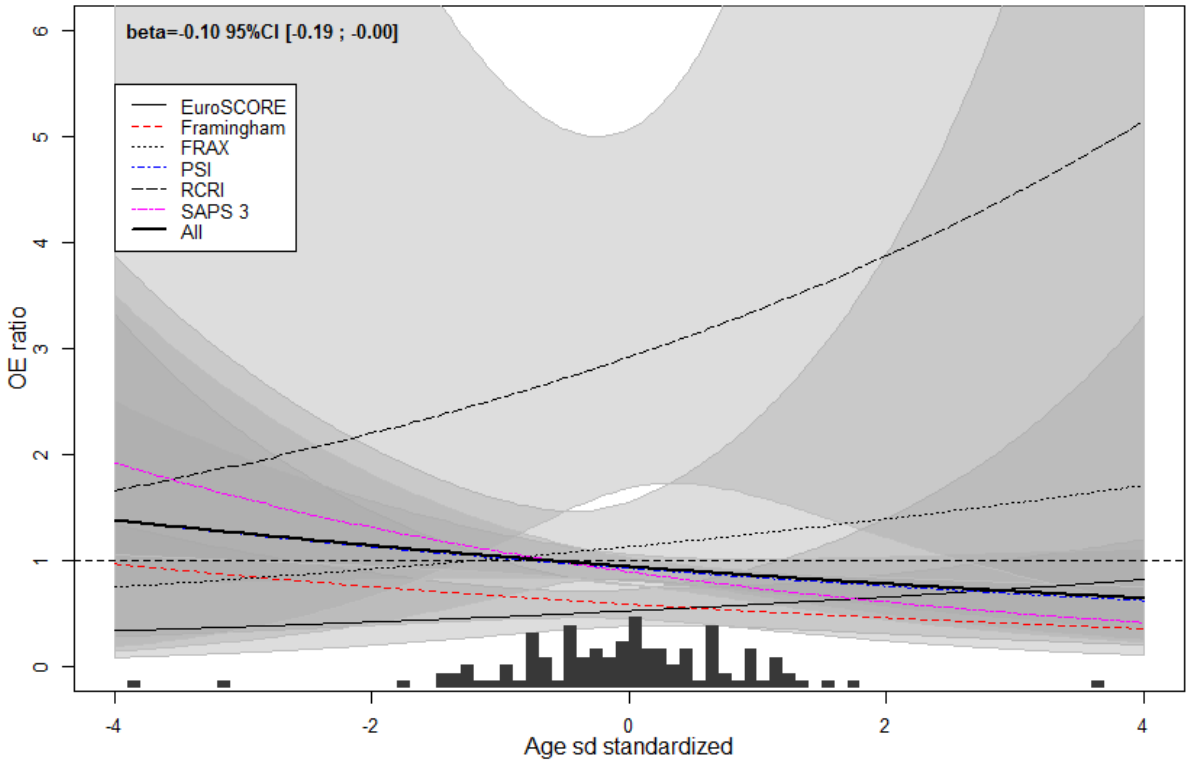

beta=-0.04 95%CI [-0.29 ; 0.21]
Framingham
PSI
RCRI
All
OE ratio
0
1
2
3
4
5
6
-6
-4
-2
0
2
4
Prediction horizon standardized


beta=0.13 95%CI [-0.02 ; 0.29]
Framingham
FRAX
PSI
RCRI
SAPS 3
All
OE ratio
0
1
2
3
4
5
6
-4
-2
0
2
4
Average follow-up duration standardized


beta=-0.06 95%CI [-0.15 ; 0.02]
EuroSCORE
Framingham
FRAX
PSI
RCRI
SAPS 3
All
OE ratio
0
1
2
3
4
5
6
-2
-1
0
1
2
3
4
5
Number of participants standardized


beta=-0.02 95%CI [-0.10 ; 0.05]
EuroSCORE
Framingham
FRAX
PSI
RCRI
SAPS 3
All
OE ratio
0
1
2
3
4
5
6
-2
-1
0
1
2
3
4
5
Number of events standardized


beta=-0.02 95%CI [-0.10 ; 0.07]
EuroSCORE
Framingham
FRAX
PSI
RCRI
SAPS 3
All
OE ratio
0
1
2
3
4
5
6
-4
-2
0
2
4
Age mean standardized


beta=-0.10 95%CI [-0.19 ; -0.00]
EuroSCORE
Framingham
FRAX
PSI
RCRI
SAPS 3
All
OE ratio
0
1
2
3
4
5
6
-4
-2
0
2
4
Age sd standardized

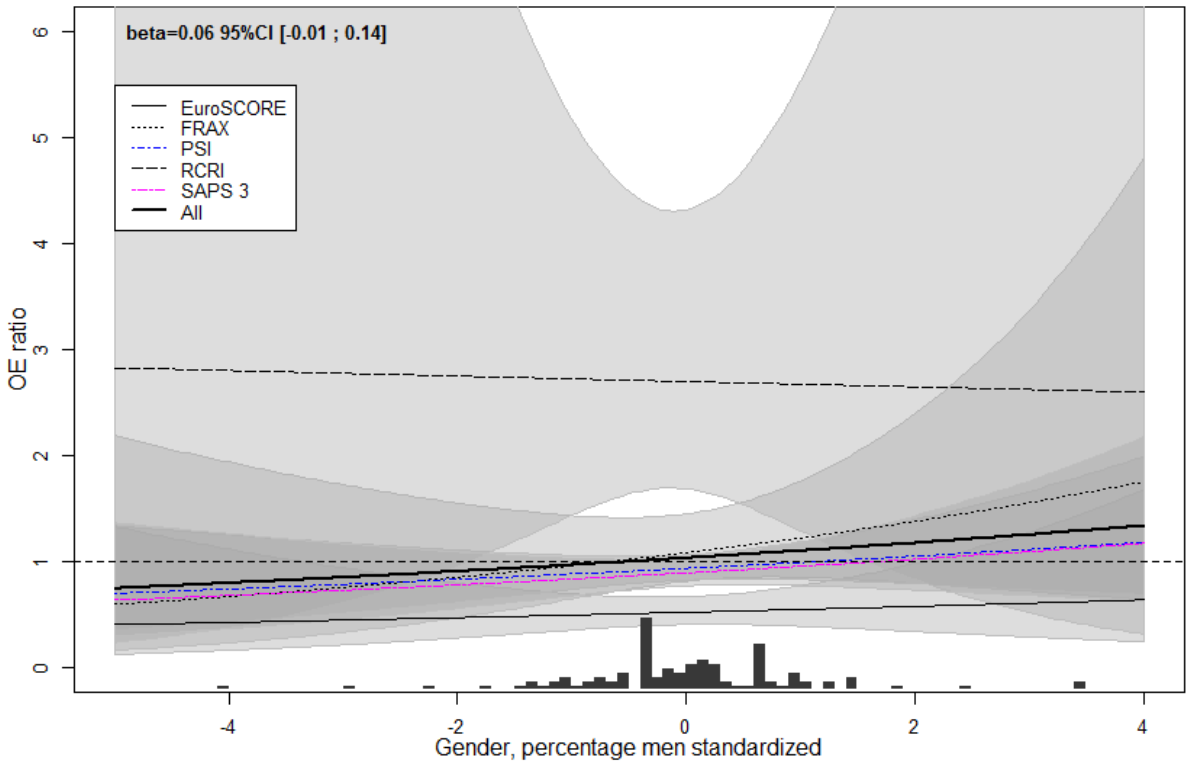
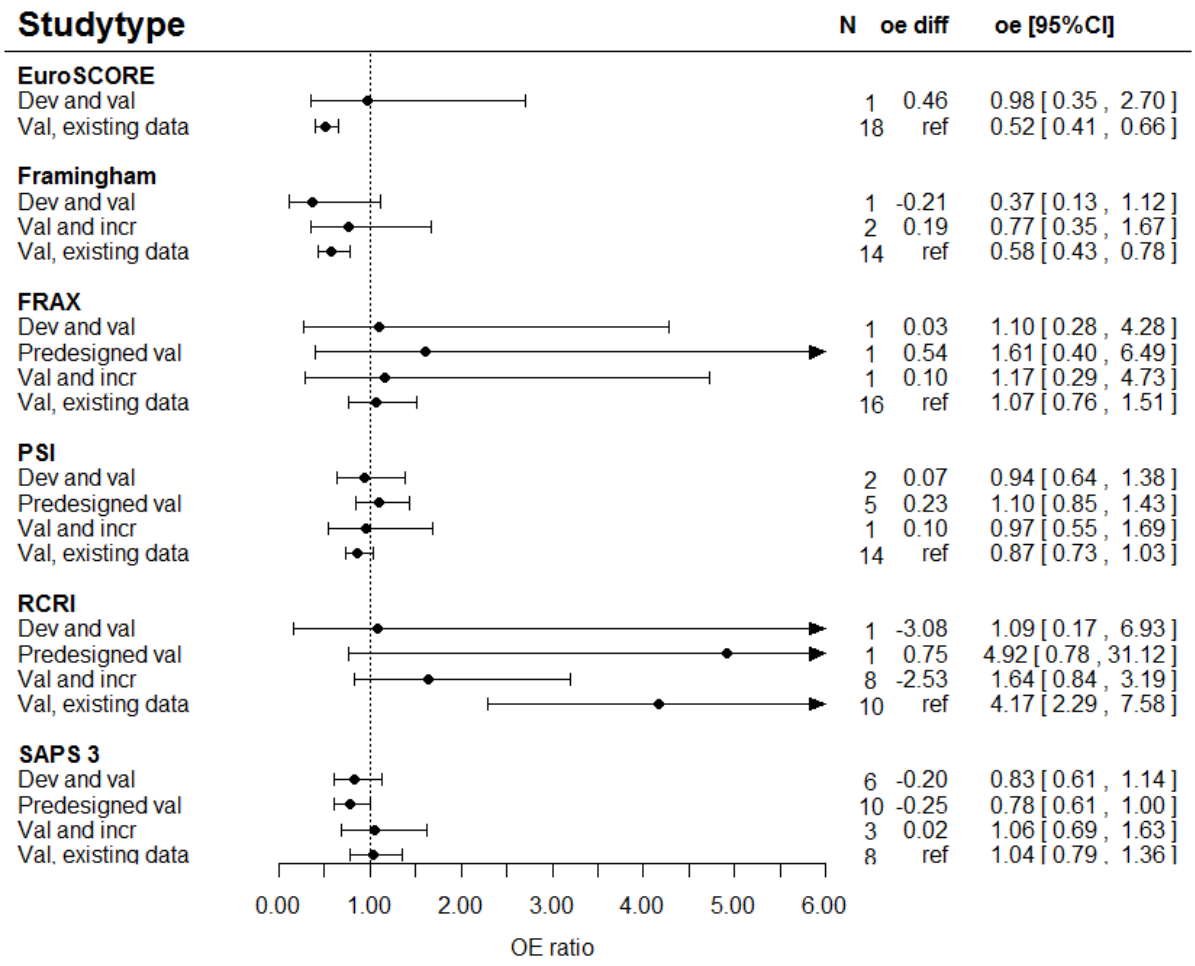
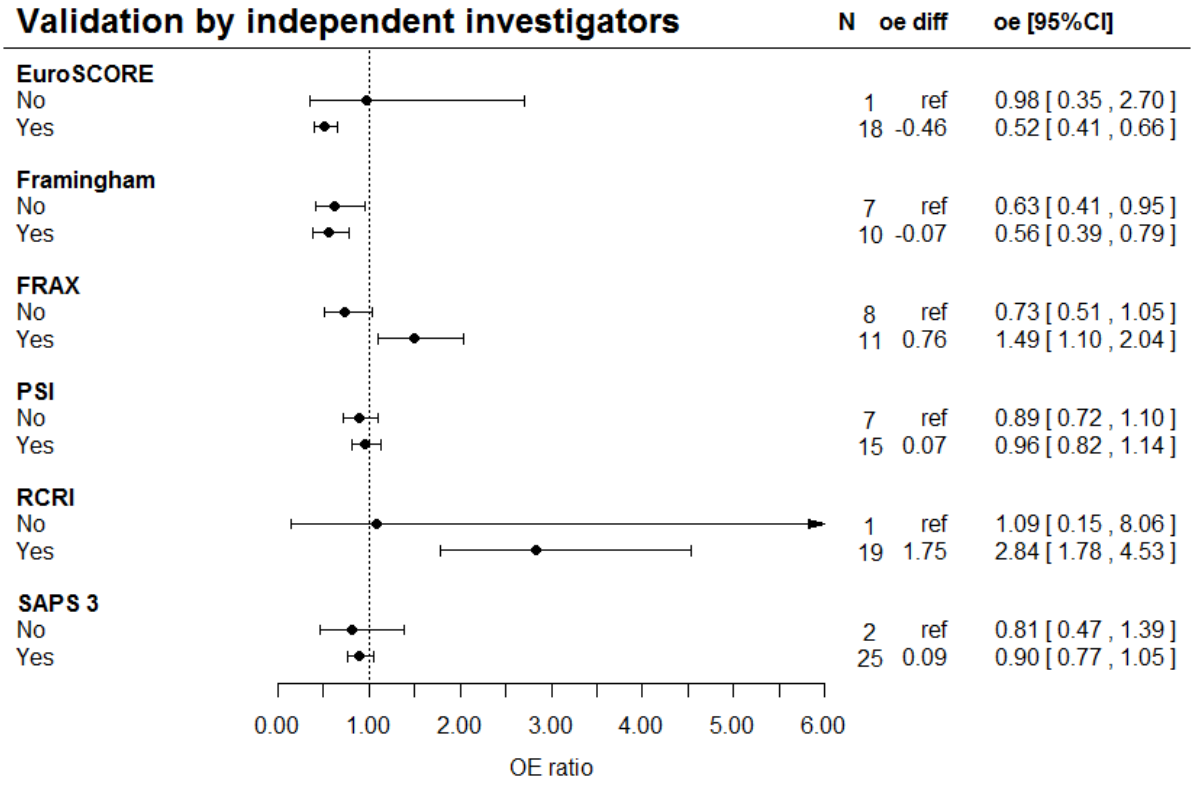
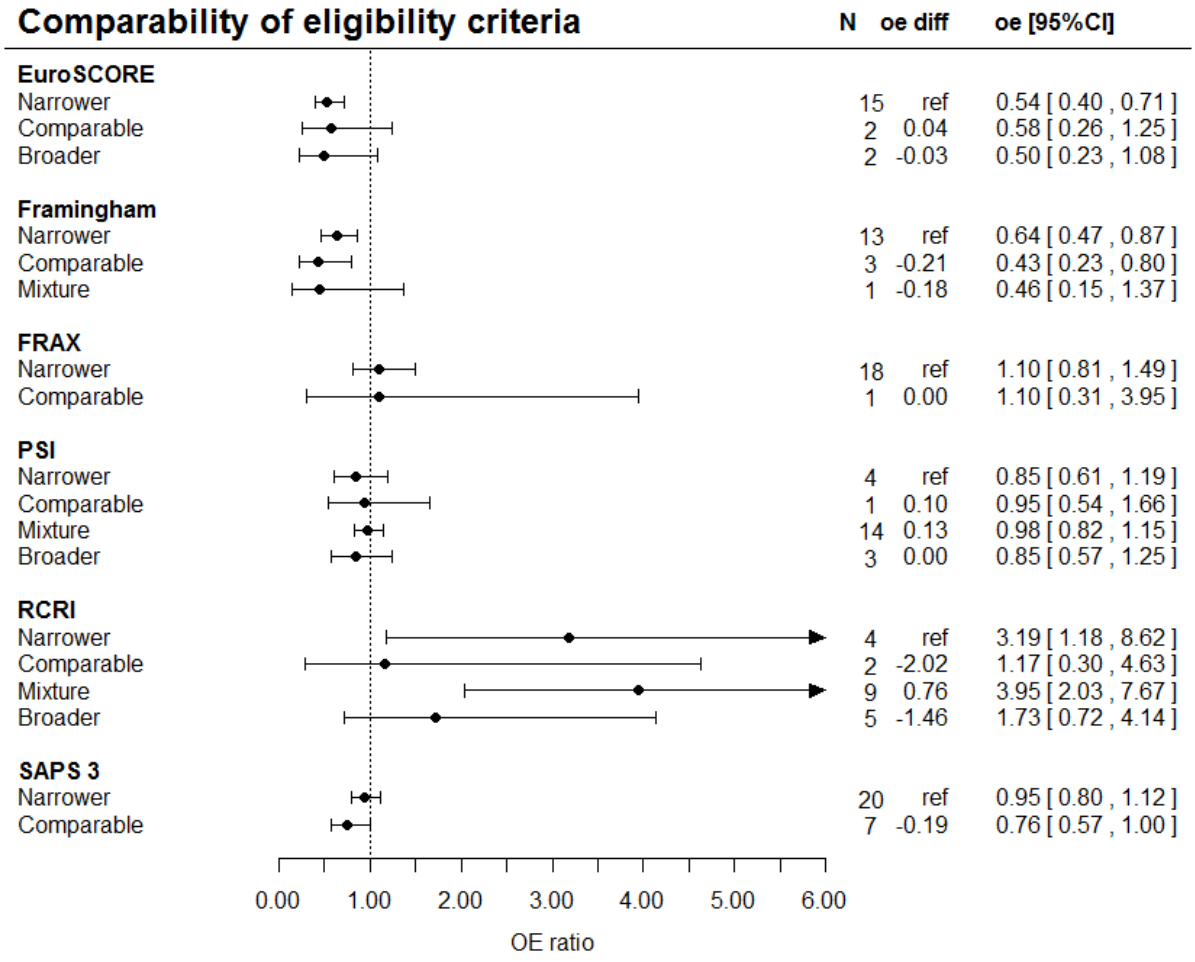
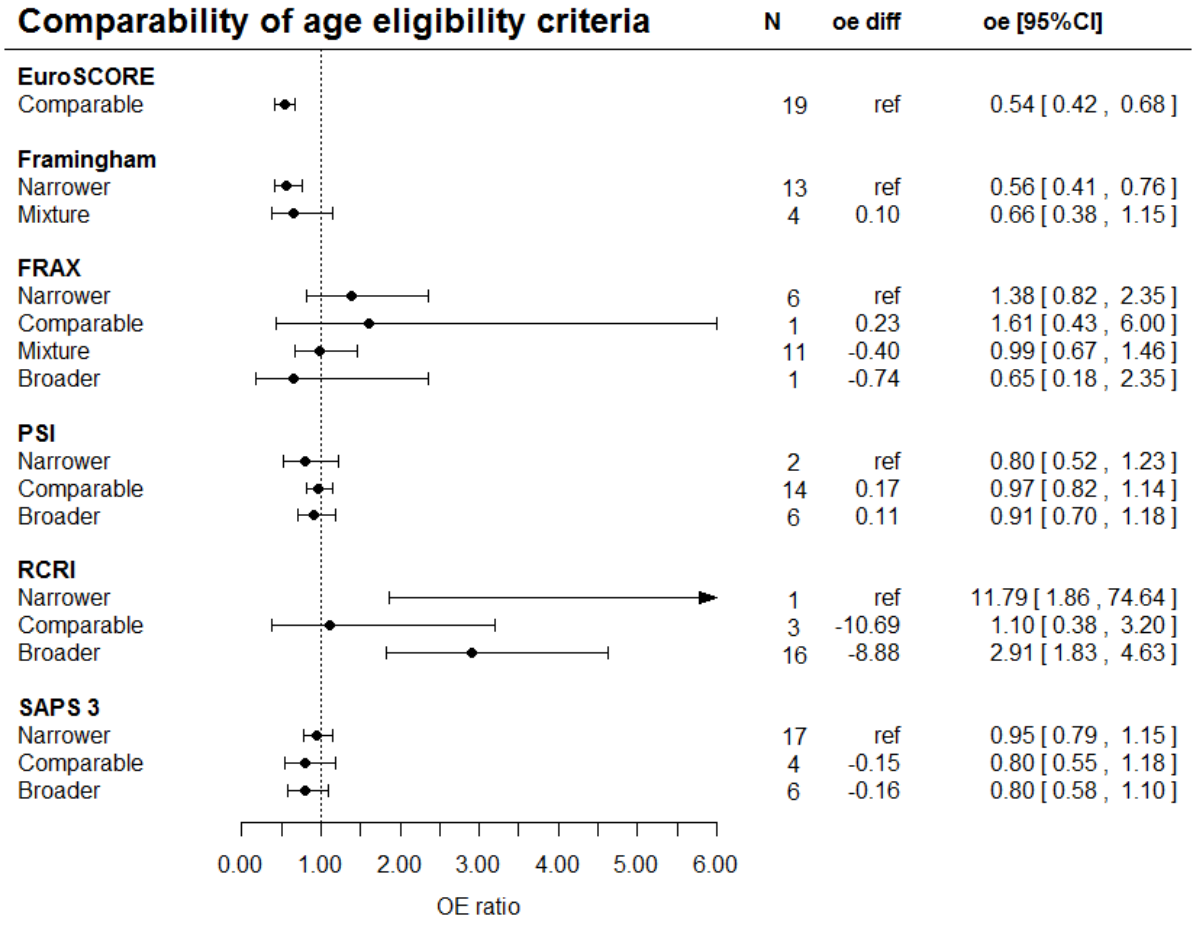

beta=0.06 95%CI [-0.01 ; 0.14]
EuroSCORE
FRAX
PSI
RCRI
SAPS 3
All
OE ratio
0
1
2
3
4
5
6
-4
-2
0
2
4
Gender, percentage men standardized

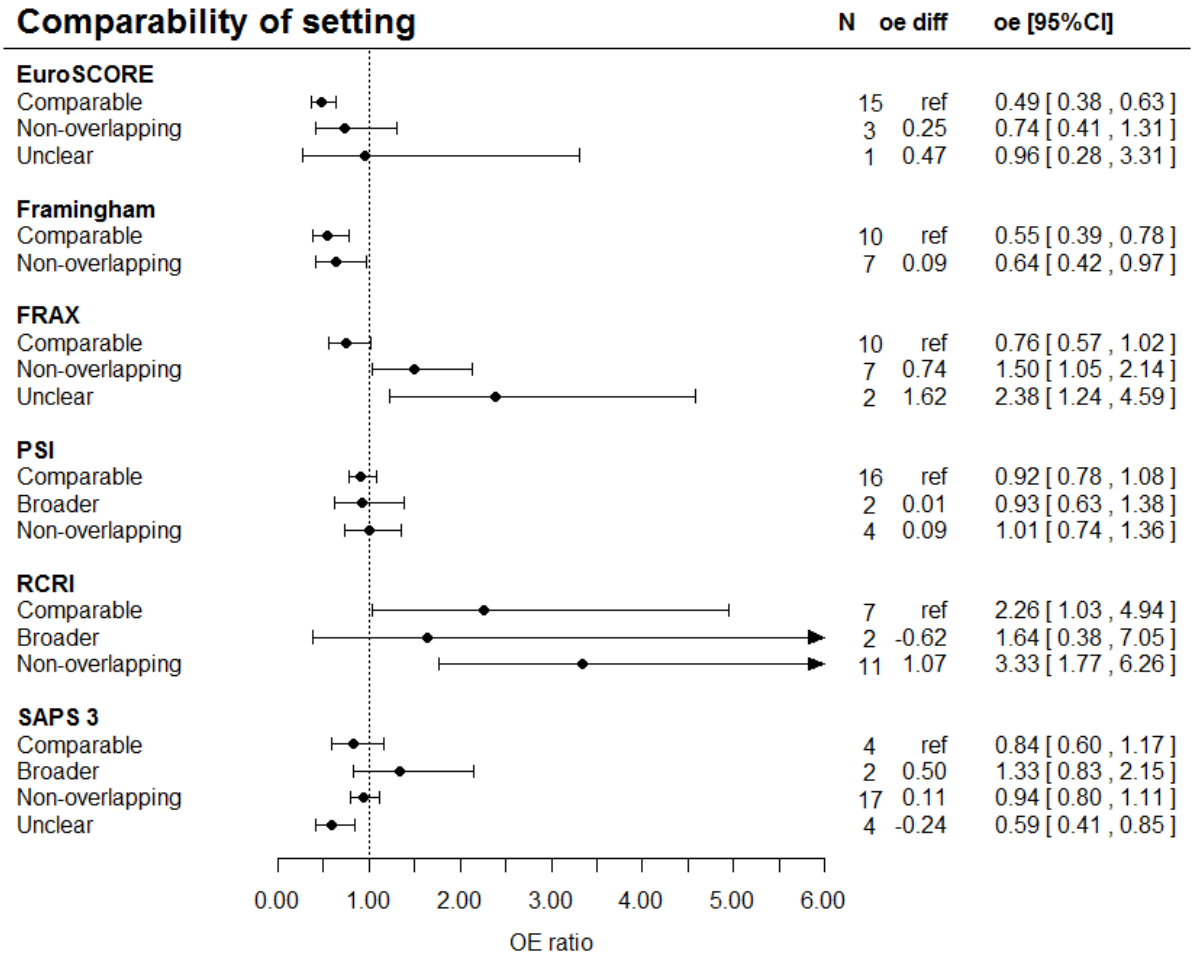
Figure S6: OE ratio in categories of study characteristics within each systematic review

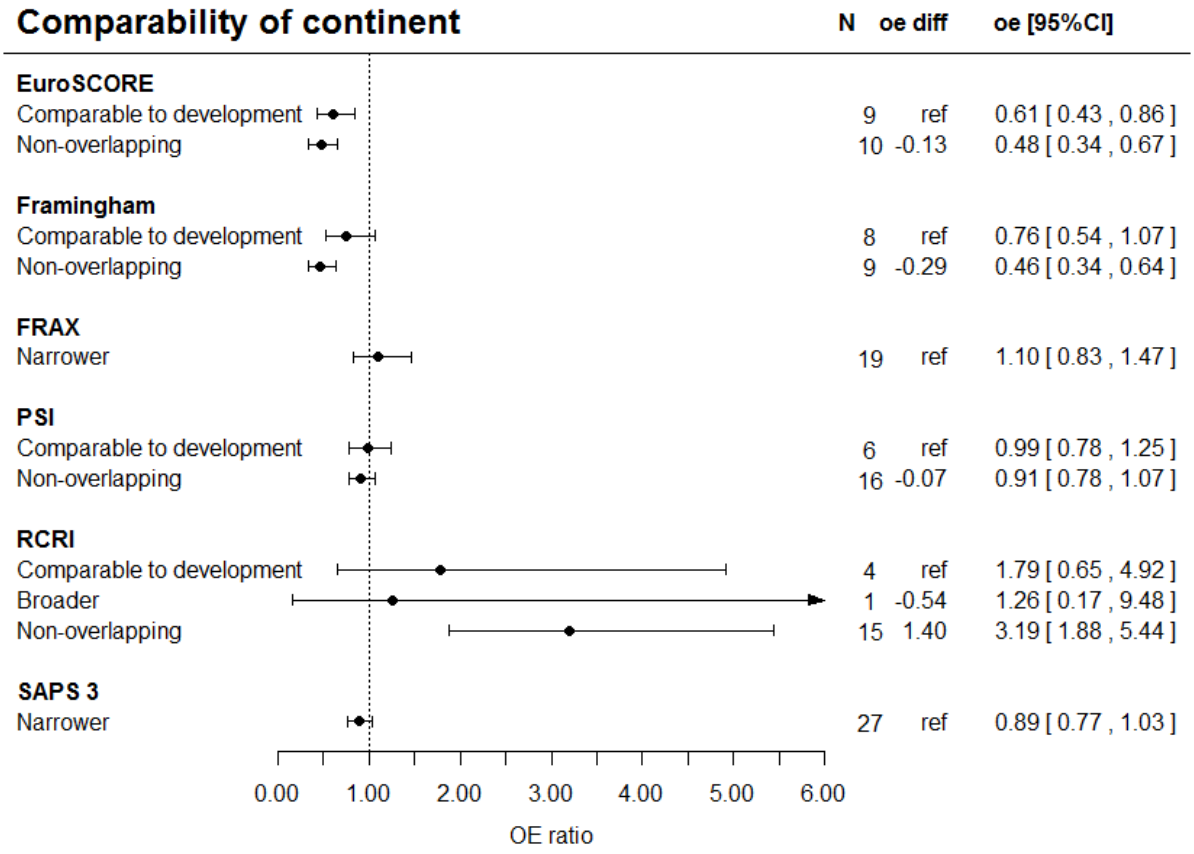


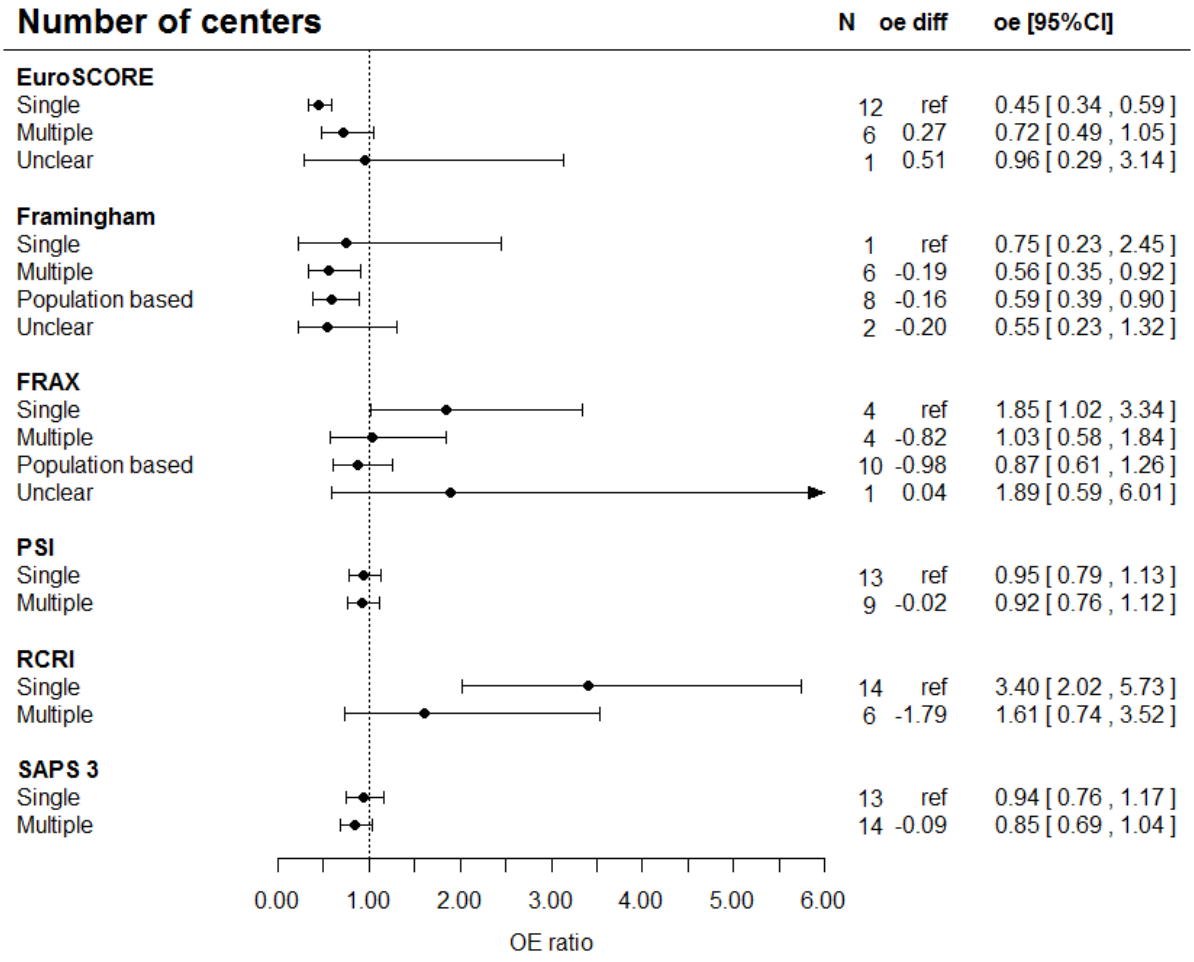


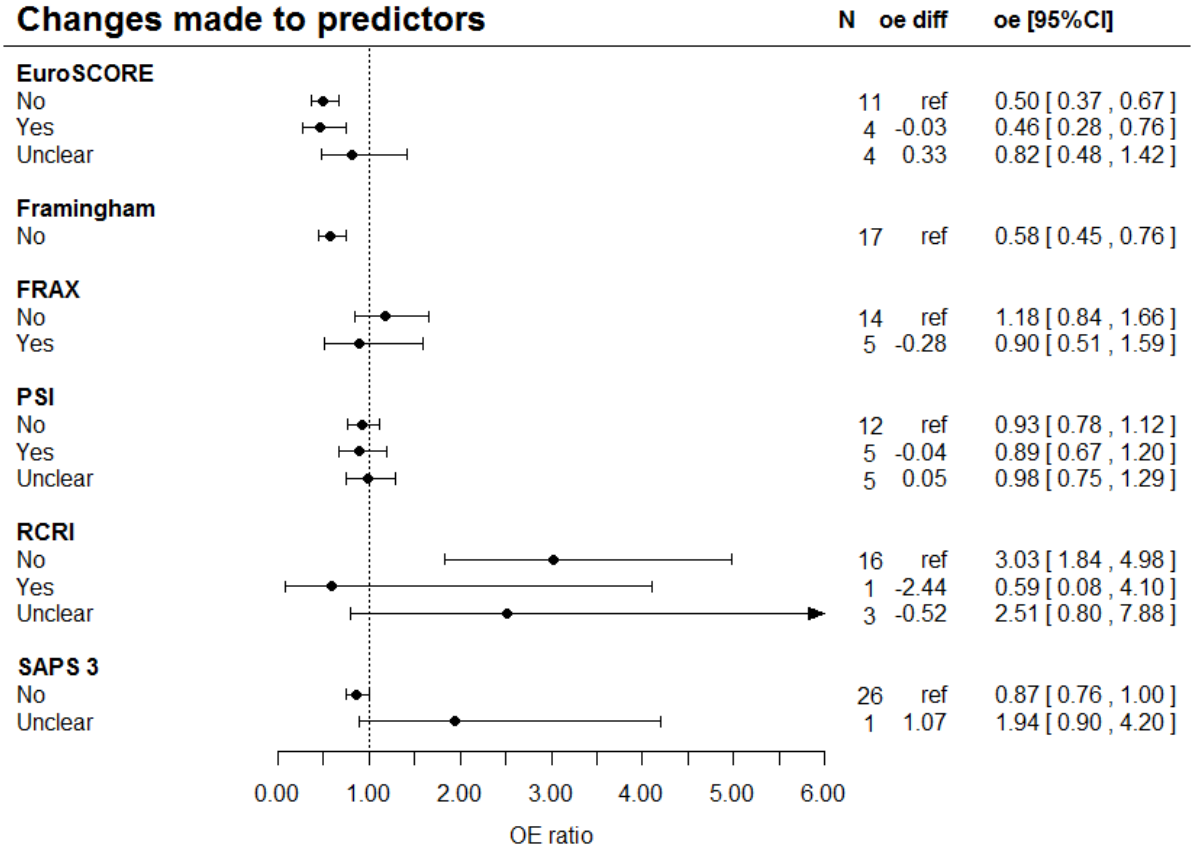

Comparability of eligibility criteria
N
oe diff
oe [95%CI]
EuroSCORE
Narrower 15 ref 0.54 [ 0.40 , 0.71 ]
Comparable 2 0.04 0.58 [ 0.26 , 1.25 ]
Broader 2 -0.03 0.50 [ 0.23 , 1.08 ]
Framingham
Narrower 13 ref 0.64 [ 0.47 , 0.87 ]
Comparable 3 -0.21 0.43 [ 0.23 , 0.80 ]
Mixture 1 -0.18 0.46 [ 0.15 , 1.37 ]
FRAX
Narrower 18 ref 1.10 [ 0.81 , 1.49 ]
Comparable 1 0.00 1.10 [ 0.31 , 3.95 ]
PSI
Narrower 4 ref 0.85 [ 0.61 , 1.19 ]
Comparable 1 0.10 0.95 [ 0.54 , 1.66 ]
Mixture 14 0.13 0.98 [ 0.82 , 1.15 ]
Broader 3 0.00 0.85 [ 0.57 , 1.25 ]
RCRI
Narrower 4 ref 3.19 [ 1.18 , 8.62 ]
Comparable 2 -2.02 1.17 [ 0.30 , 4.63 ]
Mixture 9 0.76 3.95 [ 2.03 , 7.67 ]
Broader 5 -1.46 1.73 [ 0.72 , 4.14 ]
SAPS 3
Narrower 20 ref 0.95 [ 0.80 , 1.12 ]
Comparable 7 -0.19 0.76 [ 0.57 , 1.00 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio

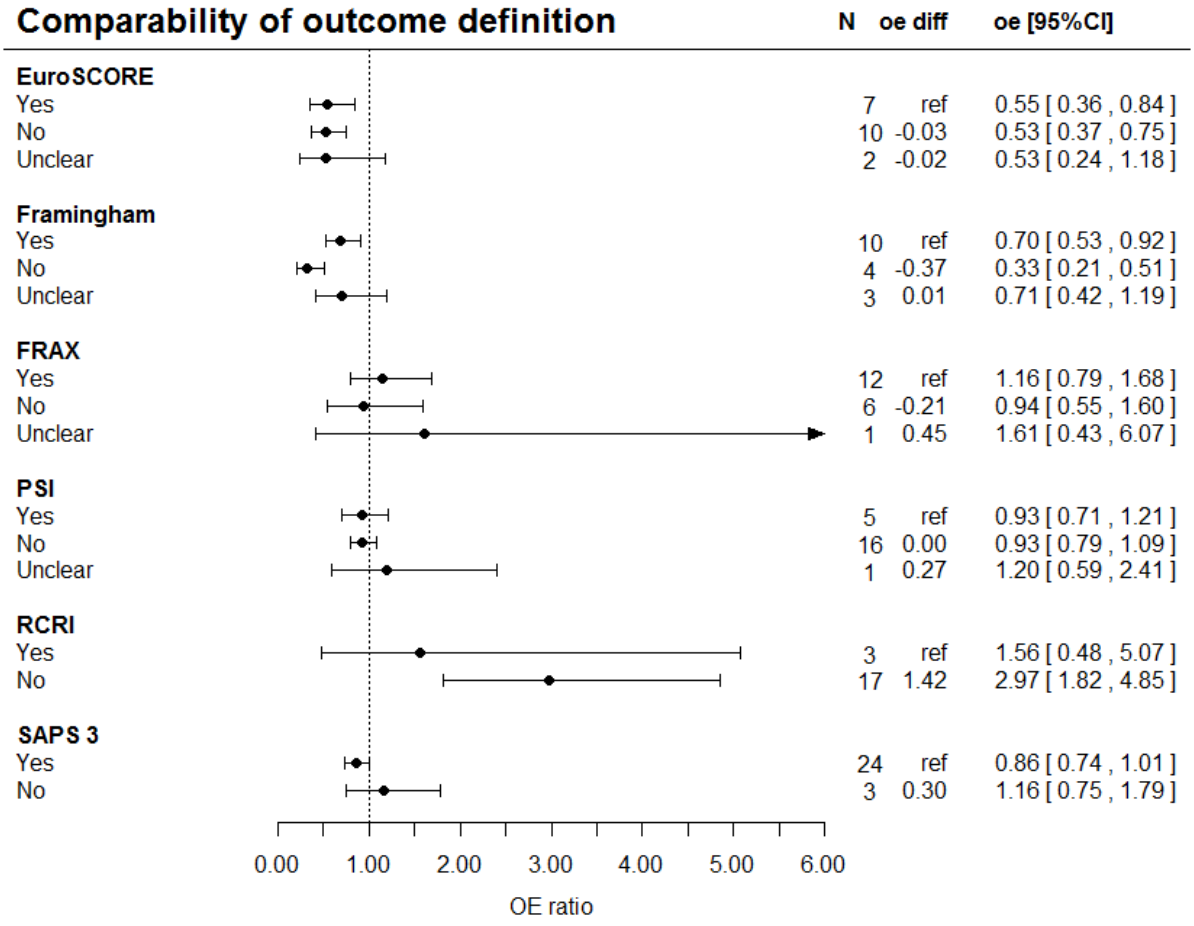

Comparability of age eligibility criteria
N
oe diff
oe [95%CI]
EuroSCORE
Comparable
19
ref
0.54 [ 0.42 , 0.68 ]
Framingham
Narrower
13
ref
0.56 [ 0.41 , 0.76 ]
Mixture
4
0.10
0.66 [ 0.38 , 1.15 ]
FRAX
Narrower
6
ref
1.38 [ 0.82 , 2.35 ]
Comparable
1
0.23
1.61 [ 0.43 , 6.00 ]
Mixture
11
-0.40
0.99 [ 0.67 , 1.46 ]
Broader
1
-0.74
0.65 [ 0.18 , 2.35 ]
PSI
Narrower
2
ref
0.80 [ 0.52 , 1.23 ]
Comparable
14
0.17
0.97 [ 0.82 , 1.14 ]
Broader
6
0.11
0.91 [ 0.70 , 1.18 ]
RCRI
Narrower
1
ref
11.79 [ 1.86 , 74.64 ]
Comparable
3
-10.69
1.10 [ 0.38 , 3.20 ]
Broader
16
-8.88
2.91 [ 1.83 , 4.63 ]
SAPS 3
Narrower
17
ref
0.95 [ 0.79 , 1.15 ]
Comparable
4
-0.15
0.80 [ 0.55 , 1.18 ]
Broader
6
-0.16
0.80 [ 0.58 , 1.10 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio

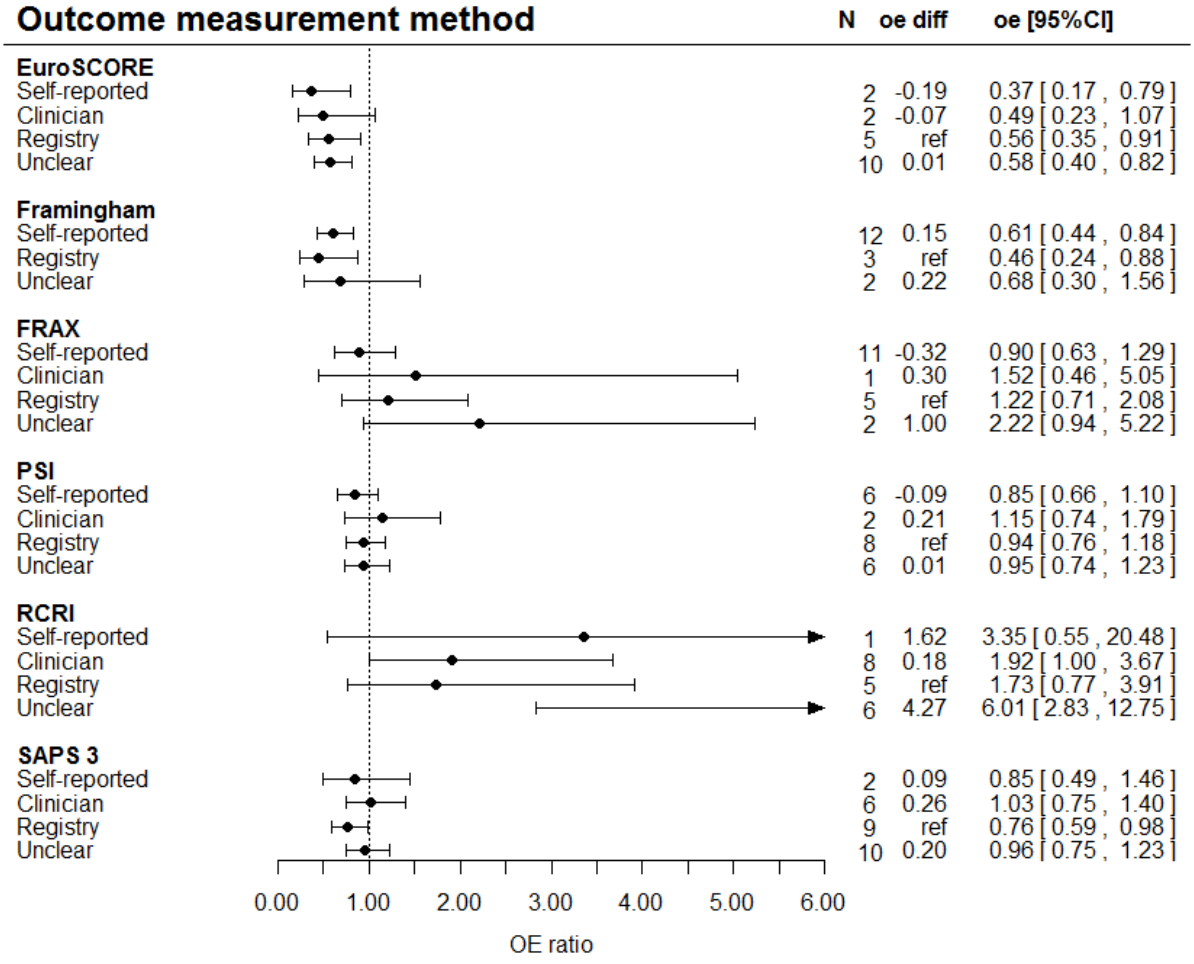

Comparability of setting
N
oe diff
oe [95%CI]
EuroSCORE
Comparable 15 ref 0.49 [ 0.38 , 0.63 ]
Non-overlapping 3 0.25 0.74 [ 0.41 , 1.31 ]
Unclear 1 0.47 0.96 [ 0.28 , 3.31 ]
Framingham
Comparable 10 ref 0.55 [ 0.39 , 0.78 ]
Non-overlapping 7 0.09 0.64 [ 0.42 , 0.97 ]
FRAX
Comparable 10 ref 0.76 [ 0.57 , 1.02 ]
Non-overlapping 7 0.74 1.50 [ 1.05 , 2.14 ]
Unclear 2 1.62 2.38 [ 1.24 , 4.59 ]
PSI
Comparable 16 ref 0.92 [ 0.78 , 1.08 ]
Broader 2 0.01 0.93 [ 0.63 , 1.38 ]
Non-overlapping 4 0.09 1.01 [ 0.74 , 1.36 ]
RCRI
Comparable 7 ref 2.26 [ 1.03 , 4.94 ]
Broader 2 -0.62 1.64 [ 0.38 , 7.05 ]
Non-overlapping 11 1.07 3.33 [ 1.77 , 6.26 ]
SAPS 3
Comparable 4 ref 0.84 [ 0.60 , 1.17 ]
Broader 2 0.50 1.33 [ 0.83 , 2.15 ]
Non-overlapping 17 0.11 0.94 [ 0.80 , 1.11 ]
Unclear 4 -0.24 0.59 [ 0.41 , 0.85 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio

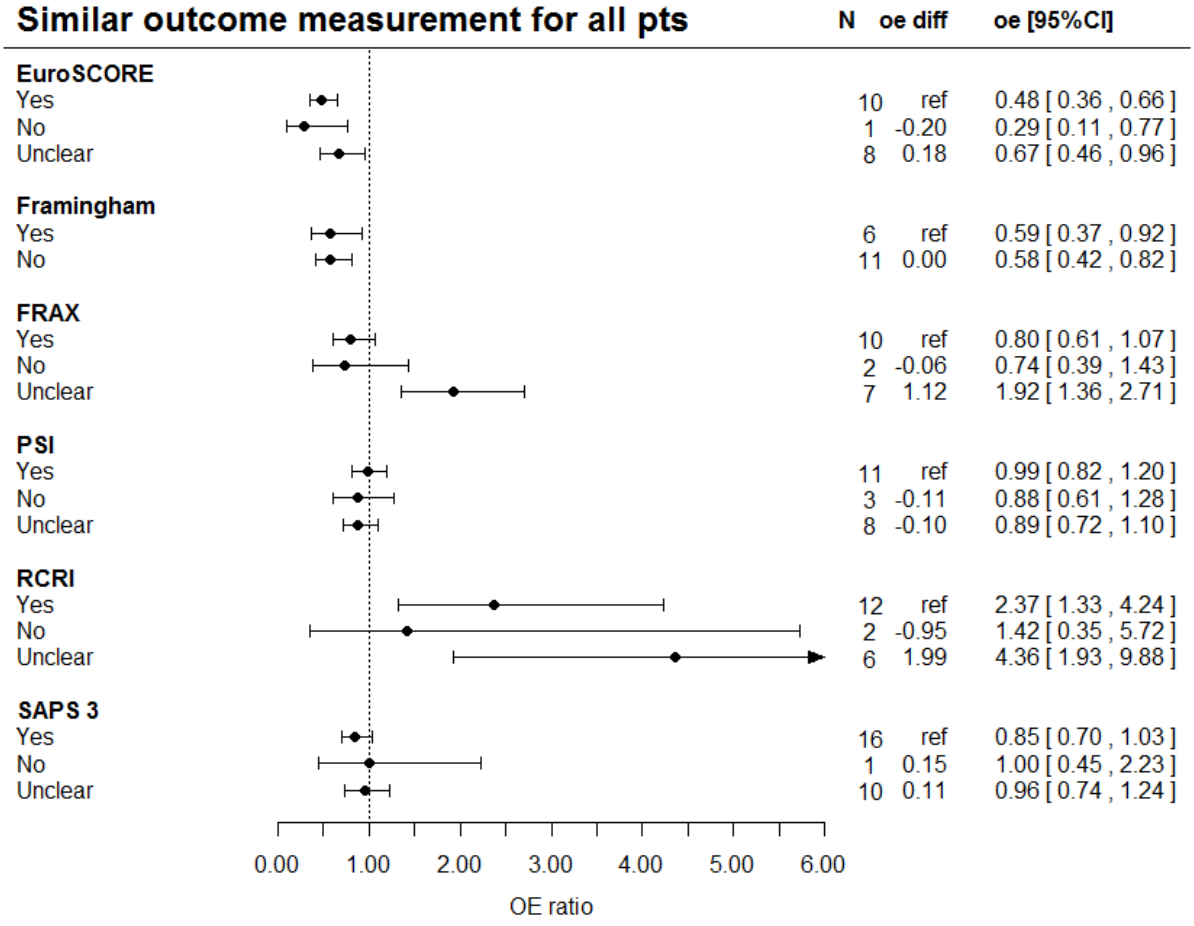

Comparability of continent
N
oe diff
oe [95%CI]
EuroSCORE
Comparable to development
9 ref 0.61 [ 0.43 , 0.86 ]
Non-overlapping
10 -0.13 0.48 [ 0.34 , 0.67 ]
Framingham
Comparable to development
8 ref 0.76 [ 0.54 , 1.07 ]
Non-overlapping
9 -0.29 0.46 [ 0.34 , 0.64 ]
FRAX
Narrower
19 ref 1.10 [ 0.83 , 1.47 ]
PSI
Comparable to development
6 ref 0.99 [ 0.78 , 1.25 ]
Non-overlapping
16 -0.07 0.91 [ 0.78 , 1.07 ]
RCRI
Comparable to development
4 ref 1.79 [ 0.65 , 4.92 ]
Broader
1 -0.54 1.26 [ 0.17 , 9.48 ]
Non-overlapping
15 1.40 3.19 [ 1.88 , 5.44 ]
SAPS 3
Narrower
27 ref 0.89 [ 0.77 , 1.03 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio

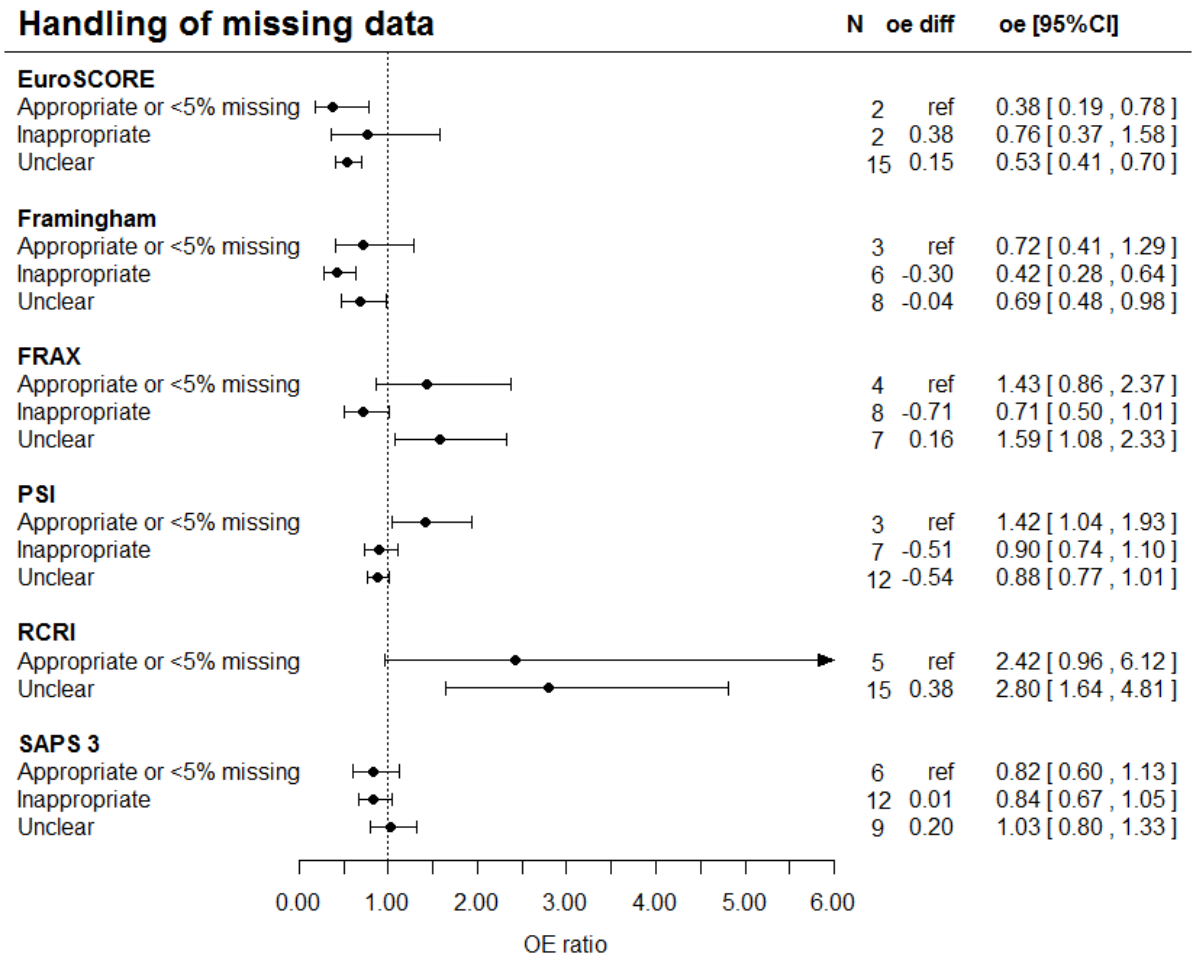

Number of centers
N
oe diff
oe [95%CI]
EuroSCORE
Single 12 ref 0.45 [ 0.34 , 0.59 ]
Multiple 6 0.27 0.72 [ 0.49 , 1.05 ]
Unclear 1 0.51 0.96 [ 0.29 , 3.14 ]
Framingham
Single 1 ref 0.75 [ 0.23 , 2.45 ]
Multiple 6 -0.19 0.56 [ 0.35 , 0.92 ]
Population based 8 -0.16 0.59 [ 0.39 , 0.90 ]
Unclear 2 -0.20 0.55 [ 0.23 , 1.32 ]
FRAX
Single 4 ref 1.85 [ 1.02 , 3.34 ]
Multiple 4 -0.82 1.03 [ 0.58 , 1.84 ]
Population based 10 -0.98 0.87 [ 0.61 , 1.26 ]
Unclear 1 0.04 1.89 [ 0.59 , 6.01 ]
PSI
Single 13 ref 0.95 [ 0.79 , 1.13 ]
Multiple 9 -0.02 0.92 [ 0.76 , 1.12 ]
RCRI
Single 14 ref 3.40 [ 2.02 , 5.73 ]
Multiple 6 -1.79 1.61 [ 0.74 , 3.52 ]
SAPS 3
Single 13 ref 0.94 [ 0.76 , 1.17 ]
Multiple 14 -0.09 0.85 [ 0.69 , 1.04 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio


Changes made to predictors
N
oe diff
oe [95%CI]
EuroSCORE
No
11 ref 0.50 [ 0.37 , 0.67 ]
Yes
4 -0.03 0.46 [ 0.28 , 0.76 ]
Unclear
4 0.33 0.82 [ 0.48 , 1.42 ]
Framingham
No
17 ref 0.58 [ 0.45 , 0.76 ]
FRAX
No
14 ref 1.18 [ 0.84 , 1.66 ]
Yes
5 -0.28 0.90 [ 0.51 , 1.59 ]
PSI
No
12 ref 0.93 [ 0.78 , 1.12 ]
Yes
5 -0.04 0.89 [ 0.67 , 1.20 ]
Unclear
5 0.05 0.98 [ 0.75 , 1.29 ]
RCRI
No
16 ref 3.03 [ 1.84 , 4.98 ]
Yes
1 -2.44 0.59 [ 0.08 , 4.10 ]
Unclear
3 -0.52 2.51 [ 0.80 , 7.88 ]
SAPS 3
No
26 ref 0.87 [ 0.76 , 1.00 ]
Unclear
1 1.07 1.94 [ 0.90 , 4.20 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio


Comparability of outcome definition
N
oe diff
oe [95%CI]
EuroSCORE
Yes 7 ref 0.55 [ 0.36 , 0.84 ]
No 10 -0.03 0.53 [ 0.37 , 0.75 ]
Unclear 2 -0.02 0.53 [ 0.24 , 1.18 ]
Framingham
Yes 10 ref 0.70 [ 0.53 , 0.92 ]
No 4 -0.37 0.33 [ 0.21 , 0.51 ]
Unclear 3 0.01 0.71 [ 0.42 , 1.19 ]
FRAX
Yes 12 ref 1.16 [ 0.79 , 1.68 ]
No 6 -0.21 0.94 [ 0.55 , 1.60 ]
Unclear 1 0.45 1.61 [ 0.43 , 6.07 ]
PSI
Yes 5 ref 0.93 [ 0.71 , 1.21 ]
No 16 0.00 0.93 [ 0.79 , 1.09 ]
Unclear 1 0.27 1.20 [ 0.59 , 2.41 ]
RCRI
Yes 3 ref 1.56 [ 0.48 , 5.07 ]
No 17 1.42 2.97 [ 1.82 , 4.85 ]
SAPS 3
Yes 24 ref 0.86 [ 0.74 , 1.01 ]
No 3 0.30 1.16 [ 0.75 , 1.79 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio


Outcome measurement method
N
oe diff
oe [95%CI]
EuroSCORE
Self-reported 2 -0.19 0.37 [ 0.17 , 0.79 ]
Clinician 2 -0.07 0.49 [ 0.23 , 1.07 ]
Registry 5 ref 0.56 [ 0.35 , 0.91 ]
Unclear 10 0.01 0.58 [ 0.40 , 0.82 ]
Framingham
Self-reported 12 0.15 0.61 [ 0.44 , 0.84 ]
Registry 3 ref 0.46 [ 0.24 , 0.88 ]
Unclear 2 0.22 0.68 [ 0.30 , 1.56 ]
FRAX
Self-reported 11 -0.32 0.90 [ 0.63 , 1.29 ]
Clinician 1 0.30 1.52 [ 0.46 , 5.05 ]
Registry 5 ref 1.22 [ 0.71 , 2.08 ]
Unclear 2 1.00 2.22 [ 0.94 , 5.22 ]
PSI
Self-reported 6 -0.09 0.85 [ 0.66 , 1.10 ]
Clinician 2 0.21 1.15 [ 0.74 , 1.79 ]
Registry 8 ref 0.94 [ 0.76 , 1.18 ]
Unclear 6 0.01 0.95 [ 0.74 , 1.23 ]
RCRI
Self-reported 1 1.62 3.35 [ 0.55 , 20.48 ]
Clinician 8 0.18 1.92 [ 1.00 , 3.67 ]
Registry 5 ref 1.73 [ 0.77 , 3.91 ]
Unclear 6 4.27 6.01 [ 2.83 , 12.75 ]
SAPS 3
Self-reported 2 0.09 0.85 [ 0.49 , 1.46 ]
Clinician 6 0.26 1.03 [ 0.75 , 1.40 ]
Registry 9 ref 0.76 [ 0.59 , 0.98 ]
Unclear 10 0.20 0.96 [ 0.75 , 1.23 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio


Similar outcome measurement for all pts
N
oe diff
oe [95%CI]
EuroSCORE
Yes 10 ref 0.48 [ 0.36 , 0.66 ]
No 1 -0.20 0.29 [ 0.11 , 0.77 ]
Unclear 8 0.18 0.67 [ 0.46 , 0.96 ]
Framingham
Yes 6 ref 0.59 [ 0.37 , 0.92 ]
No 11 0.00 0.58 [ 0.42 , 0.82 ]
FRAX
Yes 10 ref 0.80 [ 0.61 , 1.07 ]
No 2 -0.06 0.74 [ 0.39 , 1.43 ]
Unclear 7 1.12 1.92 [ 1.36 , 2.71 ]
PSI
Yes 11 ref 0.99 [ 0.82 , 1.20 ]
No 3 -0.11 0.88 [ 0.61 , 1.28 ]
Unclear 8 -0.10 0.89 [ 0.72 , 1.10 ]
RCRI
Yes 12 ref 2.37 [ 1.33 , 4.24 ]
No 2 -0.95 1.42 [ 0.35 , 5.72 ]
Unclear 6 1.99 4.36 [ 1.93 , 9.88 ]
SAPS 3
Yes 16 ref 0.85 [ 0.70 , 1.03 ]
No 1 0.15 1.00 [ 0.45 , 2.23 ]
Unclear 10 0.11 0.96 [ 0.74 , 1.24 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio




OE ratio for categories of study characteristics, pooled using univariable meta-regression analyses per systematic review. N represents the number of external validation studies in a specific category. OE diff represents the difference in OE ratio with regard to a reference category (indicated with 'ref'). Dev: development, val: validation, incr: incremental value, pts: patients.

## References


1. Snell KI, 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? *Stat Methods Med Res* 2017:962280217705678.
2. Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: asymptotic methods and evaluation. *Stat Med* 2006;25(4):559-73.
3. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29-36.
4. 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.
5. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* 2014;14:25.
6. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2015.
7. R: A language and environment for statistical computing [program]. Vienna, Austria: R Foundation for Statistical Computing, 2016.
8. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 2010;36(3):1-48.
9. Gasparrini A, Armstrong B, Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Stat Med* 2012;31(29):3821-39.
10. Debray TP. Metamisc: Diagnostic and Prognostic Meta-Analysis. 2017.
11. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw* 2015;67(1).

| Section/topic | Proposed item to be used in methodology research | Reported on page |
|---|---|---|
| **TITLE** | | |
| Title | Identify the report as a meta-epidemiologic study. | 1 |
| **ABSTRACT** | | |
| Structured summary | Provide a structured summary that includes the background of the topic, goal of the study, data sources, method of data selection, appraisal and synthesis methods, results, limitations, conclusions and implications of key findings. | 3 |
| **INTRODUCTION** | | |
| Rationale | Describe the rationale for the meta-epidemiological study in the context of what is already known. | 5 |
| Objectives | Provide an explicit statement of the goal of the meta-epidemiological study and the hypothesis being empirically tested. | 5 |
| **METHODS** | | |
| Protocol | Indicate if a protocol exists, if and where it can be accessed (eg, Web address). Registration of a protocol is not mandatory. | Available on request |
| Eligibility criteria | Specify study characteristics used as criteria for eligibility with a rationale. | 7 |
| Information sources | Describe all information sources (eg, databases with dates of coverage, contact with experts to identify additional studies, Internet searches) and search date. | 7 |
| Search | Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated. Search is commonly not driven by a clinical question. | Supplement page 2 |
| Study selection | Describe the process for selecting studies for inclusion (ie, how many reviewers selected studies, reviewing in duplicate or by single individuals). | 7,8 |
| Data collection process | Describe method of data extraction from reports (eg, piloted forms, independently, in duplicate) and any processes used for manipulating data or obtaining and confirming data from investigators. | 8,9, Supplement page 3-6 |
| Data items | List and define all variables for which data were sought and any assumptions and imputations made. | Supplement page 3-6 |
| Risk of bias in individual studies | If risk of bias assessment of individual studies was relevant to the analysis, describe the items used and how this information is to be used during data synthesis. | Not assessed |
| Summary measures | State the principal summary measures (eg, ratio of risk ratios, difference in means) and explain its meaning and direction to readers. | 9, Supplement page 7, 8 |
| Synthesis of results | Describe the statistical or descriptive methods of synthesis including measures of consistency if relevant. If applicable, describe the development of statistical or simulation modelling based on theoretical background. Describe and justify assumptions and | 9, Supplement page 7, 8 |

*From: Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. Evid Based Med 2017;22(4):139-42.*

| Section/topic | Proposed item to be used in methodology research | Reported on page |
|---|---|---|
| | computational approximations. Describe methods of additional analyses (eg, sensitivity or subgroup analyses, meta-regression), if done, indicating which were prespecified. | |
| **RESULTS** | | |
| Study selection | Give numbers of studies assessed for eligibility and included in the study, with reasons for exclusions at each stage, ideally with a flow diagram. Present a measure of inter-reviewer agreement (eg, kappa statistic). | 10, Figure 1 |
| Study characteristics | For each study, present characteristics for which data were extracted and provide the citations. Clinical characteristics may not always be relevant. | Supplement page 9-12 |
| Risk of bias within studies | If risk of bias assessment of individual studies was used in the meta-epidemiological analysis, report risk of bias indicators of each study to allow replication of findings. | Not assessed |
| Results of individual studies | Present data elements used in the meta-epidemiological analysis from each study (results of clinical outcomes may not be relevant). | Not done |
| Synthesis of results | Present results of statistical analysis done, including measures of precision and measures of consistency. Present validity of assumptions and fit of statistical or simulation modelling, if applicable. | 11, 12, Figure 2-5, Supplement page 13-49 |
| Additional analysis | Give results of additional analyses, if done (eg, sensitivity or subgroup analyses, meta-regression). | Not done |
| **DISCUSSION** | | |
| Summary of evidence | Summarise the main findings and compare them with existing knowledge about the topic. The quality of evidence may not be relevant; however, investigators should describe their certainty in the results to readers. | 13,14 |
| Limitations | Discuss limitations at research methodology level (eg, likelihood of reporting or publication bias). | 13,14 |
| Conclusions | Provide general interpretation of the results and implications for future research. Provide any plausible impact on clinical practice. | 16 |
| **FUNDING** | | |
| Funding | Describe sources of funding for the methodology research and role of funders. | 17 |

*From: Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. Evid Based Med 2017;22(4):139-42.*

BMJ Open

# Empirical evidence on the impact of study characteristics and the performance of prediction models: a meta-epidemiological study



| | |
|---|---|
| Journal: | *BMJ Open* |
| Manuscript ID | bmjopen-2018-026160.R1 |
| Article Type: | Research |
| Date Submitted by the Author: | 05-Nov-2018 |
| Complete List of Authors: | Damen, Johanna; Cochrane Netherlands, University Medical Center Utrecht, Utrecht University; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University<br>Debray, Thomas; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University; Cochrane Netherlands, University Medical Center Utrecht, Utrecht University<br>Pajouheshnia, Romin; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University<br>Reitsma, Johannes; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University; Cochrane Netherlands, University Medical Center Utrecht, Utrecht University<br>Scholten, Rob; Cochrane Netherlands, University Medical Center Utrecht, Utrecht University; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University<br>Moons, Karel; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University; Cochrane Netherlands, University Medical Center Utrecht, Utrecht University<br>Hooft, Lotty; Cochrane Netherlands, University Medical Center Utrecht, Utrecht University; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University |
| Primary Subject Heading: | Research methods |
| Secondary Subject Heading: | Epidemiology |
| Keywords: | Meta-epidemiology, Prognosis, Prognostic models, Bias, Prediction |
| | |

# Empirical evidence on the impact of study characteristics and the performance of prediction models: a meta-epidemiological study

Johanna A A G Damen, Thomas P A Debray, Romin Pajouheshnia, Johannes B Reitsma, Rob J P M Scholten, Karel G M Moons, Lotty Hooft

Johanna A A G Damen
Assistant professor
Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Thomas P A Debray
Assistant professor
Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Romin Pajouheshnia
Postdoctoral researcher
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Johannes B Reitsma
Associate professor
Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Rob J P M Scholten
Professor

Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Karel G M Moons
Professor
Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Lotty Hooft
Associate professor
Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, 3508 GA Utrecht, The Netherlands

Correspondence to:
Johanna A A G Damen
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht
P.O. Box 85500
Str. 6.131
3508 GA Utrecht
The Netherlands
j.a.a.damen@umcutrecht.nl
+31 88 75 693 77

Word count: 4167

## Abstract

**Objectives:** To empirically assess the relation between study characteristics and prognostic model performance in external validation studies of multivariable prognostic models.

**Design:** Meta-epidemiological study.

**Data sources and study selection:** On October 16th, 2018, we searched electronic databases for systematic reviews of prognostic models. Reviews from non-overlapping clinical fields were selected if they reported common performance measures (either the concordance (c)-statistic or the ratio of observed over expected number of events (OE ratio)) from ten or more validations of the same prognostic model.

**Data extraction and analyses:** Study design features, population characteristics, methods of predictor and outcome assessment, and the aforementioned performance measures were extracted from the included external validation studies. Random effects meta-regression was used to quantify the association between the study characteristics and model performance.

**Results:** We included 10 systematic reviews, describing a total of 224 external validations, of which 221 reported c-statistics and 124 OE ratios. Associations between study characteristics and model performance were heterogeneous across systematic reviews. C-statistics were most associated with variation in population characteristics, outcome definitions and measurement, and predictor substitution. For example, validations with eligibility criteria comparable to the development study were associated with higher c-statistics compared to narrower criteria (difference in logit c-statistic 0.21 [95% CI 0.07, 0.35], similar to an increase from 0.70 to 0.74). Using a case-control design was associated with higher OE ratios, compared to using data from a cohort (difference in log OE ratio 0.97 [95% CI 0.38, 1.55], similar to an increase in OE ratio from 1.00 to 2.63).

**Conclusions:** Variation in performance of prognostic models across studies is mainly associated with variation in case-mix, study designs, outcome definitions and measurement methods, and predictor substitution. Researchers developing and validating prognostic models should realise the potential influence of these study characteristics on the predictive performance of prognostic models.

**Strengths and limitations of this study**

- To the best of our knowledge, this is the first meta-epidemiological study focusing on the association of study characteristics with estimates of prognostic model performance.
- We included all ten systematic reviews describing at least ten external validations of the same prognostic model, resulting in 224 external validations.
- We extracted relevant features of design and conduct according to existing checklists on quality assessment (CHARMS) and reporting of prediction model studies (TRIPOD).
- It was not feasible to fit multivariable meta-regression models due to the limited number of available, well-reported, validation studies within the individual reviews, rendering the effective sample size too small for multivariable meta-regression analyses.

## Introduction

Prediction models, including diagnostic and prognostic models, estimate the probability that an individual has or will develop a certain outcome (e.g. disease or complication). Hereto, they combine multiple predictors into an estimate of an individual's risk.[1] Before using a prediction model in clinical practice it is recommended to validate the performance of the model in a population other than the population in which the model was developed (so called external validation studies).[2] Such studies assess whether model predictions remain sufficiently accurate across different settings and populations. Obviously, it is important that the methodological quality of external validation studies is good, as otherwise estimates of the prediction model's performance may be biased and thereby lead to misleading conclusions on its generalizability to practice.

Systematic reviews have found that the performance of existing prediction models often varies substantially across external validation studies of those models.[3-5] These differences may not only appear due to random variation (when validation studies are small), but may also arise when model predictions are invalid because the model is applied in very different populations (eg, the association between predictors in the model and the outcome are different) or when design-related characteristics of the validation study (eg, measurement methods or variable definitions) are not well aligned with the original development study.[2 6]

To provide empirical evidence of the association of study characteristics with prediction model performance, a meta-epidemiological approach can be used. Studies using this approach have shown the influence of study characteristics on the effectiveness of interventions studied in randomized trials and on the accuracy of diagnostic tests.[7-12] For diagnostic prediction models evidence suggests estimates of performance may be biased in studies with certain study characteristics. One study found a higher diagnostic odds ratio in case-control studies, studies with differential outcome verification (ie, using different outcome assessments across study individuals), and with low sample size.[13] To date, no meta-epidemiological study has been performed investigating the possible impact of study characteristics on measures of the predictive performance of a prognostic model upon external validation, which is commonly quantified in terms of discrimination and calibration.[14] The aim of this study was to investigate sources of heterogeneity in the predictive performance of prognostic models. A meta-

epidemiological approach was used to synthesize evidence from a range of clinical fields. This study can serve as empirical evidence for design and analysis related bias in prognostic model studies.

## Methods

Search and selection of systematic reviews

We used an existing database (last updated on October 16th, 2018) consisting of studies evaluating multiple existing prediction models, including narrative or systematic reviews of prediction models, or head-to-head comparisons of multiple prediction models validated on a specific dataset (See Supplement 1 for details of the search strategy and selection criteria). To construct this database, references identified by the search were screened for eligibility by one reviewer (GSC) on title, abstract and, if necessary, on full text. Subsequently, the full text of all articles in the database were screened for eligibility to the current project by another reviewer (JAAGD). We selected systematic reviews of prognostic models (ie, diagnostic models were excluded) that included at least ten studies that externally validated the same prognostic model (this number was chosen a priori to facilitate the estimation of study-level parameters such as between-study heterogeneity), and that presented the performance of these models in terms of discrimination (concordance (c)-statistic or area under the receiver operating characteristic (AUC) curve), or calibration (observed expected (OE) ratio). Discrimination is the ability of the model to distinguish between people who will and who will not develop the outcome of interest, while calibration reflects the overall agreement between the total number of observed and predicted ('expected') events.[14] We excluded systematic reviews that selected studies based on specific study characteristics (eg, we excluded systematic reviews that did not include primary studies with a sample size below 100, if we were not able to identify the primary studies that had been excluded for this reason). Furthermore, we excluded reviews of prognostic models in which the weights of predictors in the original model were based on expert opinion rather than on coefficients estimated from a formal statistical approach. If more than one systematic review on the same prognostic model was identified, we included the one with the broadest inclusion criteria (eg, reviews focussing on specific patient populations were not preferred if a review with a broader population was available) or the most recent review (in this order of preference). When multiple prognostic models for the same condition were described in one systematic review which all fulfilled the selection criteria, we included the model with the highest number of external validations.

Selection of the primary external validation studies from the included systematic reviews

From the included systematic reviews we collected the primary studies in which the prognostic models were developed and externally validated. For primary external validation studies for which no measure

of discrimination (c-statistic) or calibration (total OE ratio) was reported in the systematic review, we checked the full text of the primary study, and if performance was not reported, these studies were excluded.

If primary external validation studies described multiple external validations of the same model and if there was no overlap in included participants between these external validations (eg, a model was validated in two different cohorts, or a model was validated in men and women separately), data were extracted for every external validation separately. If a model was validated multiple times on the same population (described in either one or multiple publications), we selected the external validation that was included in the systematic review. If the systematic review included all those external validations, we selected the one in which the study population and predicted outcome most closely resembled the population and outcome of the original model.

Data extraction and preparation

We extracted relevant features of design and conduct according to existing checklists on quality assessment (CHARMS) and reporting of prediction model studies (TRIPOD).[15-17] Information about study characteristics of studies in which the models were developed were extracted from the corresponding development papers. Information about study characteristics of primary external validation studies were first extracted from systematic reviews. This information was subsequently checked using the external validation studies and, if necessary, additional information was extracted by one reviewer (JAAGD or RP). Items we extracted included study type (eg, external validation only, development of a new model and external validation of a model), study design (eg, existing cohort, existing RCT), dependency of investigators (validation by independent investigators or investigators also involved in the development study), eligibility criteria for participant inclusion, setting, location (continent), study dates, number of centres, follow-up time and prediction horizon, age and gender distribution, deletion or substitution of predictors, outcome definition and measurement method, sample size and number of events, handling of missing data, and model performance (see Supplement 2 for details). The data extraction form was piloted on multiple articles by all reviewers (JAAGD, TPAD, LH, KGMM, RP, JBR, RJPMS).

For analysis purposes, some study characteristics had to be categorized or transformed (Supplement 2). For example, eligibility criteria of the validation study as compared to the development study had to be judged and categorized as comparable, narrower (if subgroups included in the development study were excluded from the validation study), broader (if subgroups excluded from the development study were

included in the validation study), mixture (a combination of the two), or unclear. For setting, location, predictors and outcome a similar categorization was used. If data on study characteristics were not reported in the primary external validation studies, these were either categorized as 'unclear' (in case of categorical study variables), or the study was excluded from the analyses of that (missing) study characteristic (in case of continuous study variables, such as sample size). In order to improve comparability between reviews, we standardized continuous study variables separately for each systematic review, i.e. for every variable we subtracted the mean and divided by the standard deviation of all external validations identified from the same systematic review.

Statistical analyses

We used a two-staged approach to study the possible association between study characteristics and predictive performance.

In the first stage, we fitted a univariable meta-regression model for every study characteristic within each systematic review with the logit c-statistic or log OE ratio as outcome variable.[18] The regression coefficients estimated from this meta-regression model indicate the difference in logit c-statistic or log OE ratio between a certain category of a study characteristic and a chosen reference category (ie, the category that was present in most systematic reviews) of that characteristic.

In the second stage, these regression coefficients were pooled by the use of a random effects model. This reflected the average influence of the study characteristic on model performance across all systematic reviews. For continuous characteristics, the regression coefficients obtained in the first stage were jointly pooled across reviews, using bivariate meta-analysis.[19 20] For categorical characteristics the results of univariable meta-analyses are presented. We planned to perform multivariable analyses to assess the association between various study characteristics in combination and the performance of prognostic models, but due to the paucity of data we were not able to do so. All analyses are described in more detail in Supplement 3.

Patient and public involvement

Patients and public were not involved in the design, recruitment or conduct of the study.

## Results

Identification and selection of studies

The search identified 2392 studies, of which 555 were included in the database and screened on full text, and 79 were further assessed (Figure 1). Finally, ten systematic reviews were included.[21-29] These reviews addressed external validations of the following prognostic models: ABCD2,[30] Essen Stroke Risk Score (ESRS),[31] EuroSCORE,[32] Framingham,[33] FRAX,[34] Injury Severity Score (ISS),[35] model for end-stage liver disease (MELD),[36] Pneumonia Severity Index (PSI),[37] Revised Cardiac Risk Index (RCRI),[38] and Simplified Acute Physiology Score (SAPS) 3[39] (Table 1). The reviews included 248 primary external validation studies with 274 external model validations (one study could describe multiple model validations). During data extraction, 73 of 274 validations were eventually excluded (most often for not reporting a performance measure), and 20 additional external model validations were identified (Figure 1). This resulted in the inclusion of 224 external validations, of which 221 could be included in the analyses of the c-statistic, and 124 in the analyses of the total OE ratio. For the total OE ratio, only validations of the EuroSCORE, Framingham, FRAX, PSI, RCRI and SAPS 3 prognostic models were included, due to the very low number of reported OE ratios in the validations studies for the other four prognostic models.

Description of included validations

The number of external validations per systematic review ranged from 11 to 30 (Table 1), and the median (IQR) sample size and number of events were 1069 (418-3043) and 92 (36-248), respectively. Most studies used an existing registry (N=104, 46%) or existing cohort (N=74, 33%) to validate the prognostic model. The median (IQR) c-statistic and total OE ratio were 0.73 (0.64-0.82) and 0.92 (0.64-1.26), respectively. Predictive performance of the models was highly heterogeneous, even for external validations of the same prognostic model, as indicated by the wide prediction intervals (Table 1).
Not all information on the study characteristics was reported for all external validations (Table S1). Information was often unclear (eg, for outcome definitions (N=83, 37%) and handling of missing data (N=105, 47%)) or missing (eg, case-mix information such as mean age (N=28, 13%) and gender distribution (N=16, 7%)).

Discrimination

*Pooled models*

The pooled analyses across all systematic reviews (Figure 2, S1 and S2) showed that validation in a continent different from the development study was associated with a higher c-statistic, compared to validation in the same continent, and multicentre versus single centre validation studies were associated with a lower c-statistic. Comparable eligibility criteria for participant inclusion were also associated with higher c-statistics compared to narrower criteria, whereas a broader setting was associated with a lower c-statistic compared to a setting comparable to the development study. Although not statistically significant, validations with changes made to the predictors (ie, substitution or deletion of a predictor), or in which it was unclear whether all predictors were correctly measured, tended to have lower c-statistics compared to validations where no changes were made. In various reviews we found an association between the c-statistic and numerous other study characteristics, such as the study design, comparability of outcome definition, prediction horizon, sample size and number of events, and mean age of study participants (Figure 3, S2 and S3), only these were often not statistically significant when pooled together.

*Variation across reviews*

Across reviews we found associations of many study characteristics with the c-statistic although this was rather heterogeneous, and confidence intervals often overlapped (Figure 3 and Figure S3). For example, for study design, in six systematic reviews a higher c-statistic was found for validations that used an existing registry compared to an existing cohort, while in three reviews a lower c-statistic was found. In three systematic reviews we found a higher c-statistic in validations by independent investigators, while in five a lower c-statistic was found.

For other study characteristics, directions of associations were more consistent. For example, for most systematic reviews, validation studies with eligibility criteria narrower compared to the criteria used in the development study had a lower c-statistics while broader eligibility criteria were associated with higher c-statistics (Figure S3). C-statistics were also (slightly) higher in external validations with a setting comparable to the development study. Validation in a continent other than the development study in general was associated with a higher c-statistic, and multicentre studies had lower c-statistics compared to single centre studies. External validations in which it was unclear if there were changes made to the predictors had lower c-statistics (Figure S3).

Calibration

*Pooled analyses*

We found a significant association between study design and the total OE ratio (Figure 4); using data from a case-control study (although known to be an inferior design for prognostic model research[1 6]) resulted in higher OE ratios, compared to using data from an existing cohort (though based on three external validations). Furthermore, higher OE ratios were found for studies in which the outcome was assessed by a panel of clinicians as compared to using a registry. In various reviews we found an association between the total OE ratio and numerous other study characteristics, such as the duration of follow-up, year in which recruitment was started, sample size, standard deviation of age, and setting (Figure 4, S4, S5 and S6), only these were not statistically significant when pooled together.

*Variation across reviews*

For other categories of study design (other than the use of a case-control design), heterogeneous associations were found across systematic reviews (Figure 5). The associations of most other study characteristics with the OE ratio were also most often not consistent across systematic reviews (Figure S5 and S6). For example, for two systematic reviews external validations with appropriate handling of missing data had OE ratios closer to 1 compared to inappropriate handling of missing data, while in two reviews, OE ratios were further away from 1. Only for the continent in which the model was validated, directions were more consistent; OE ratios were closer to 1 if the continent was comparable to the development, compared to validations in different continents (Figure S6).

## Discussion

### Principal findings

Using a comprehensive meta-analytical approach, we studied the association between study characteristics of prognostic model validation studies and the estimated model performance across ten clinical domains. We focused on objective study characteristics that can be extracted from published reports. The reporting of the primary external validation studies was often incomplete and inadequate. Key study characteristics, such as outcome definitions, handling of missing data, and even model calibration estimates were infrequently reported. Still, we found associations between various study characteristics and a model's predictive performance. Changes in a model's predictive performance were notably found in relation to validation studies with a case-control (versus cohort) design, with differences in case-mix, in continent (in which the model is validated), in eligibility criteria, in clinical setting, in number of centres (included in the validation study), in differences in outcome definitions and assessments, and in predictor substitutions.

### Explanations, strengths and weaknesses

Based on findings in meta-epidemiological studies on the effect of study characteristics and the efficacy of interventions[7-10] and diagnostic test accuracy,[11 12] we anticipated to find more statistically significant associations between study characteristics and model performance across the included systematic reviews from different domains. Although we included every systematic review that described at least ten external validation studies of the same prognostic model, resulting in more than 200 validations from 10 reviews, our analyses appeared to still be hampered by relatively low numbers of external validations per systematic review, combined with poor reporting and substantial heterogeneity within and across systematic reviews.

Conceptually, there are many potential sources of heterogeneity in model performance, such as differences in population characteristics, predictor and outcome definitions and measurements, and in many aspects of the statistical analyses (eg, dealing with missing data, sample size and selective loss to follow up). All these characteristics may act in isolation but could also be related to each other. The individual strength of the association of one characteristic with model performance is ideally addressed by adopting multivariable (meta)-regression models with the observed model performance estimates of the validation studies as dependent variable and the characteristics of multiple design features as independent variables.[10 12] Unfortunately, this approach was not feasible here due to the limited

number of available, well-reported, validation studies within the individual reviews, rendering the effective sample size too small for multivariable meta-regression analyses.

A general limitation of all meta-epidemiological studies, is the possibility that the effect of a certain study characteristic differs across systematic reviews which may nullify the effect when pooled together.[40] We also found numerous conflicting associations between a study characteristic and the reported predictive performance measures across reviews that were cancelled out in the pooled analyses.

Also, it is possible that the effect caused by individual study characteristics is small and therefore difficult to detect. Moreover, there might be some misclassification of study characteristics, caused either by our misinterpretation of what is reported, or by a lack of reporting, which could have diluted the effects of the study characteristics. Indeed, the c-statistic is often considered to be an insensitive measure to quantify changes in model performance.[41-43] In previous simulation studies, the c-statistic and OE ratio appeared to be strongly influenced by case-mix differences,[14 44 45] which may mask the possible (smaller) effects from design-related characteristics. Other measures that are less sensitive to case-mix differences, such as the calibration slope, could, however, not be studied here simply because they were (almost) never reported in our retrieved studies, as was also shown previously.[3]

We found greater variation in the methods used by external validation studies *between* models than within validations of the *same* model. For example, multiple imputation is the preferred method for handling missing data in prediction modelling.[46 47] However, in the field of cardiovascular disease, it seems common to handle missing data by performing a complete case analysis, while in the field of mortality prediction in surgical patients, typically researchers fill in 'normal' values if a value is missing. Finally, given the explorative nature of our analyses to identify potential areas of further research, we did not correct for multiple testing, though we tried to minimize the number of exploratory analyses.

Comparison to previous research

Despite above considerations, our findings, ie, the trends in the associations between study characteristics and model performance measures (though not always statistically significant), are in agreement with various previous simulation studies in this field.[14 44 46-48] For example, we confirmed that studies with more variation in case-mix show higher c-statistics, and lower c-statistics when a predictor was omitted from the model. However, we found lower c-statistics in studies with a broader setting and when the number of centres in a study was higher. The lower c-statistic in multicentre studies might be caused by increased variation in predictor definitions and methods to measure predictors compared to

single centre studies, where it is more likely that definitions and measurement methods have been standardized. This might result in increased measurement error in multicentre studies, which is known to lower discriminative ability of a model.[49 50]

We also found a higher total OE ratio in studies with a case-control design. Both simulation studies and meta-epidemiological studies in the fields of diagnostic tests and (mainly diagnostic) prediction models, have shown biased effect measures in studies using a case-control design.[11-14] This confirms that case-control studies should not be used to study certain aspects of model calibration. Further, we found that the total OE ratio was influenced by the method of outcome assessment, in agreement with previous studies that showed that higher diagnostic odds ratios were found in studies with differential outcome verification.[13] We also expected to find lower OE ratios when the validation population differed from the development population (eg, in terms of case-mix).[14] We could not systematically confirm this across all reviews, likely caused by heterogeneity between systematic reviews as indicated by the wide confidence intervals. Finally, we could not fully confirm the association between sample size and model performance that was previously found,[13] although we found similar trends in part of the reviews.

Implications for future research

In agreement with many previously conducted systematic reviews on prediction models,[3 51-55] we still and again found poor reporting of prediction model studies. Meta-epidemiological studies of prediction model studies would highly benefit from complete reporting according to the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement.[16 17] We recommend at least the following information, essential for comparing and interpreting the results of external validation studies, to be reported by every external validation study: eligibility criteria for participant inclusion, details of predictor and outcome definitions and measurements, a clear reference to the model that is being validated and any changes made to this validated model compared to the model as presented in the development study, estimates of model discrimination and calibration performance (including calibration slope and intercept), and corresponding standard errors for the original model and, if applicable, for any updated model.

We also believe that more research is urgently needed to evaluate under which circumstances certain design choices may lead to heterogeneity in prediction model performance, and to incorporate these issues in the appraisal of prediction model studies. There is a need for more guidance on how to score items of critical appraisal checklists for prediction model studies, such as the CHARMS checklist.[15]

Several options exist to gain more empirical insight in design related bias in prediction model studies. Firstly, meta-epidemiological researchers can collect more external validation studies and try to correct for all issues that cause variation in performance of a model. We believe, however, that this is currently not feasible as we already included every systematic reviews describing at least ten validations of the same prognostic model. A second and much more efficient option is to collect the individual participant data (IPD) for all studies included in this review to directly study the effect of study characteristics on model performance.[56-60] Using IPD, it will also be possible to study different performance measures, like the case-mix adjusted c-statistic[44 61] and calibration slope.[14] Thirdly, new simulation studies could be performed to get more insight in design related bias in prediction model performance. Researchers could for example study the effect of using a different outcome definition or prediction horizon on the c-statistic of a model.

**Conclusion**

In this comprehensive meta-epidemiological study we found empirical evidence for an association between study characteristics and predictive performance of prognostic models. We found that predictive performance of prognostic models upon external validation is highly heterogeneous, but sensitive to various study characteristics, such as study design, case-mix, eligibility criteria, setting, methods of outcome definition and measurement, and predictor substitution. It is important that these characteristics are thus emphasized in the reporting and appraisal of prediction model studies. However, for a large part the observed heterogeneity in model performance remained unexplained, which is likely caused by the high number of factors that cause heterogeneity in predictive performance and may act in opposite directions whereas a multivariable meta-regression analysis across reviews simply was not possible.

Acknowledgments: The authors would like to acknowledge Prof. Gary S Collins (GSC) for building the database with systematic reviews of prediction models, which served as a basis for this paper.

Contributors: KGMM, JBR, TPAD, and LH conceived the study. JAAGD, TPAD, RP, JBR, RJPMS, KGMM and LH were involved in designing the study. JAAGD selected the articles. JAAGD and RP extracted the data. JAAGD analysed the data in close consultation with TPAD. JAAGD, TPAD, RP, JBR, RJPMS, KGMM and LH were involved in interpreting the data. JAAGD wrote the first draft of the manuscript which was revised by TPAD, RP, JBR, RJPMS, KGMM and LH. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. JAAGD is guarantor.

Funding: Thomas Debray gratefully acknowledges the Netherlands Organization for Health Research and Development (grant number 91617050). Karel GM Moons received a grant from The Netherlands Organization for Scientific Research (ZONMW 918.10.615 and 91208004). The funder had no role in the design of the study; the collection, analysis, and interpretation of the data; or approval of the finished manuscript.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no 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; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Data sharing: no additional data are available.

Transparency: The lead author (JAAGD) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

## References


1. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012;98(9):683-90.
2. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2014.
3. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;353:i2416.
4. Alba AC, Agoritsas T, Jankowski M, Courvoisier D, Walter SD, Guyatt GH, et al. Risk prediction models for mortality in ambulatory patients with heart failure: a systematic review. *Circ Heart Fail* 2013;6(5):881-9.
5. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. *Breast Cancer Res Treat* 2012;132(2):365-77.
6. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ* 2009;338:b375.
7. Page MJ, Higgins JP, Clayton G, Sterne JA, Hrobjartsson A, Savovic J. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. *PLoS One* 2016;11(7):e0159267.
8. Berkman ND, Santaguida PL, Viswanathan M, Morton SC. AHRQ Methods for Effective Health Care. *The Empirical Evidence of Bias in Trials Measuring Treatment Differences*. Rockville (MD): Agency for Healthcare Research and Quality (US), 2014.
9. Savovic J, Jones H, Altman D, Harris R, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. *Health Technol Assess* 2012;16(35):1-82.
10. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 2008;336(7644):601-5.
11. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282(11):1061-6.
12. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174(4):469-76.
13. Ban JW, Emparanza JI, Urreta I, Burls A. Design Characteristics Influence Performance of Clinical Prediction Rules in Validation: A Meta-Epidemiological Study. *PLoS One* 2016;11(1):e0145779.
14. Steyerberg E. *Clinical prediction models: a practical approach to development, validation, and updating*: Springer Science & Business Media, 2008.
15. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11(10):e1001744.
16. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162(1):55-63.
17. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.

18. Snell KI, 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? *Stat Methods Med Res* 2017:962280217705678.
19. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2015.
20. 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.
21. Thompson DD, Murray GD, Dennis M, Sudlow CL, Whiteley WN. Formal and informal prediction of recurrent stroke and myocardial infarction after stroke: a systematic review and evaluation of clinical prediction models in a new cohort. *BMC Med* 2014;12:58.
22. Siregar S, Groenwold RH, de Heer F, Bots ML, van der Graaf Y, van Herwerden LA. Performance of the original EuroSCORE. *Eur J Cardiothorac Surg* 2012;41(4):746-54.
23. Marques A, Ferreira RJ, Santos E, Loza E, Carmona L, da Silva JA. The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis. *Ann Rheum Dis* 2015;74(11):1958-67.
24. Tohira H, Jacobs I, Mountain D, Gibson N, Yeo A. Systematic review of predictive performance of injury severity scoring tools. *Scand J Trauma Resusc Emerg Med* 2012;20:63.
25. Klein KB, Stafinski TD, Menon D. Predicting survival after liver transplantation based on pre-transplant MELD score: a systematic review of the literature. *PLoS One* 2013;8(12):e80661.
26. Chalmers JD, Mandal P, Singanayagam A, Akram AR, Choudhury G, Short PM, et al. Severity assessment tools to guide ICU admission in community-acquired pneumonia: systematic review and meta-analysis. *Intensive Care Med* 2011;37(9):1409-20.
27. Ford MK, Beattie WS, Wijeysundera DN. Systematic review: prediction of perioperative cardiac complications and mortality by the revised cardiac risk index. *Ann Intern Med* 2010;152(1):26-35.
28. Nassar AP, Malbouisson LM, Moreno R. Evaluation of Simplified Acute Physiology Score 3 performance: a systematic review of external validation studies. *Crit Care* 2014;18(3):R117.
29. Damen JAAG, Pajouheshnia R, Heus P, Moons KGM, Reitsma JB, Scholten RJPM, et al. Performance of the Framingham risk models and Pooled Cohort Equations for predicting 10-year risk of cardiovascular disease: a systematic review and meta-analysis. Manuscript submitted for publication.
30. Rothwell PM, Giles MF, Flossmann E, Lovelock CE, Redgrave JN, Warlow CP, et al. A simple score (ABCD) to identify individuals at high early risk of stroke after transient ischaemic attack. *Lancet* 2005;366(9479):29-36.
31. Diener HC, Ringleb PA, Savi P. Clopidogrel for the secondary prevention of stroke. *Expert Opin Pharmacother* 2005;6(5):755-64.
32. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg* 1999;16(1):9-13.
33. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97(18):1837-47.
34. Kanis JA, Oden A, Johnell O, Johansson H, De Laet C, Brown J, et al. The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. *Osteoporos Int* 2007;18(8):1033-46.
35. Baker SP, O'Neill B, Haddon W, Jr., Long WB. The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. *J Trauma* 1974;14(3):187-96.

36. Malinchoc M, Kamath PS, Gordon FD, Peine CJ, Rank J, ter Borg PC. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology* 2000;31(4):864-71.
37. Fine MJ, Auble TE, Yealy DM, Hanusa BH, Weissfeld LA, Singer DE, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med* 1997;336(4):243-50.
38. Lee TH, Marcantonio ER, Mangione CM, Thomas EJ, Polanczyk CA, Cook EF, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation* 1999;100(10):1043-9.
39. Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P, Campos RA, et al. SAPS 3--From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* 2005;31(10):1345-55.
40. Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med* 2002;21(11):1513-24.
41. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21(1):128-38.
42. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., 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(2):157-72; discussion 207-12.
43. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011;30(10):1105-17.
44. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010;172(8):971-80.
45. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. *BMJ* 2016;353:i3139.
46. Held U, Kessels A, Garcia Aymerich J, Basagana X, Ter Riet G, Moons KG, et al. Methods for Handling Missing Variables in Risk Prediction Models. *Am J Epidemiol* 2016;184(7):545-51.
47. Janssen KJ, Vergouwe Y, Donders AR, Harrell FE, Jr., Chen Q, Grobbee DE, et al. Dealing with missing predictor values when applying clinical prediction models. *Clin Chem* 2009;55(5):994-1001.
48. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol* 2012;12:82.
49. Khudyakov P, Gorfine M, Zucker D, Spiegelman D. The impact of covariate measurement error on risk prediction. *Stat Med* 2015;34(15):2353-67.
50. Pajouheshnia R, van Smeden M, Peelen LM, Groenwold RHH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *J Clin Epidemiol* 2018.
51. 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.
52. 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(3):268-77.
53. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9(5):1-12.
54. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak* 2006;6:38.

55. 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.
56. Debray TP, Koffijberg H, Vergouwe Y, Moons KG, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. *Stat Med* 2012;31(23):2697-712.
57. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* 2013;32(18):3158-80.
58. Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS Med* 2015;12(10):e1001886.
59. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, 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.
60. Riley RD, Price MJ, Jackson D, Wardle M, Gueyffier F, Wang J, et al. Multivariate meta-analysis using individual participant data. *Res Synth Methods* 2015;6(2):157-74.
61. White IR, Rapsomaniki E. Covariate-adjusted measures of discrimination for survival data. *Biom J* 2015;57(4):592-613.
62. Giles MF, Rothwell PM. Systematic review and pooled analysis of published and unpublished validations of the ABCD and ABCD2 transient ischemic attack risk scores. *Stroke* 2010;41(4):667-73.

Table 1: Description of included reviews and prediction models

| **Systematic review** | **Giles 2010[62]** | **Thompson 2014[21]** | **Siregar 2012[22]** | **Damen[29]** | **Marques 2015[23]** | **Tohira 2012[24]** | **Klein 2013[25]** | **Chalmers 2011[26]** | **Ford 2010[27]** | **Nassar 2014[28]** |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | ABCD2[30] | ESRS[31] | EuroSCORE[32] | Framingham[33] | FRAX[34] | ISS[35] | MELD[36] | PSI[37] | RCRI[38] | SAPS 3[39] |
| Population | Patients with TIA | Adults with a previous CVD event | Adult patients who underwent cardiac surgery under cardiopulmonary bypass | Men without previous CHD event | General population | Injured patients | Patients with liver cirrhosis but without hepatocellular carcinoma who underwent elective transjugular intrahepatic portosystemic shunts | Inpatients with community-acquired pneumonia | Patients aged >=50 years who underwent nonemergent noncardiac procedures | ICU patients |
| Geographical location (continent) | United States and UK (Europe and North America) | Canada, United States, Europe (Europe and North America) | Europe (Europe) | United States (North America) | Europe, Canada, Japan, United States, Australia (Europe, North America, Asia, Australia) | United States (North America) | United States (North America) | United States (North America) | United States (North America) | Worldwide (all continents) |
| Patient recruitment | 1981-1998 | 1992-1995 | 1995 | 1971-1974 | 1980-1999 | 1968-1969 | 1991-1995 | 1989 | 1989-1994 | 2002 |
| Predicted outcome | Stroke | Recurrent ischemic stroke, MI and vascular death | Mortality | CHD | Osteoporotic fractures | All-cause mortality | All-cause mortality | 30-day hospital mortality | Major cardiac complications | Hospital mortality |
| Prediction horizon | 2 days | 1 year | 30 days | 10 years | 10 years | 3 months | 3 months | 30 days | 1 year | 90 days |
| **Performance development study** | | | | | | | | | | |
| C-statistic | 0.66 [95% CI 0.60, 0.71] | NR | 0.7875 | 0.74 | 0.63 | NR | NR | 0.84 | 0.759 [SE 0.032] | 0.848 |
| OE ratio | NR* | NR* | NR* | NR* | NR* | NR* | NR* | NR* | NR* | 1.00 [95% CI 0.98, 1.02] |
| ***Pooled* performance validation studies** | | | | | | | | | | |
| Number of external validations included in analyses | 16 | 11 | 22 | 23 | 30 | 34 | 14 | 24 | 23 | 27 |
| C-statistic [95% CI] | 0.66 | 0.60 | 0.79 | 0.68 | 0.66 | 0.86 | 0.64 | 0.80 | 0.69 | 0.83 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | [0.61, 0.71] | [0.58, 0.62] | [0.77, 0.81] | [0.65, 0.71] | [0.63, 0.68] | [0.83, 0.88] | [0.59, 0.68] | [0.77, 0.82] | [0.65, 0.72] | [0.80, 0.85] |
| 95% PI | [0.54, 0.77] | [0.57, 0.63] | [0.74, 0.83] | [0.56, 0.78] | [0.54, 0.76] | [0.62, 0.96] | [0.48, 0.77] | [0.64, 0.89] | [0.53, 0.81] | [0.66, 0.92] |
| OE ratio [95% CI] | NA | NA | 0.54 [0.42, 0.68] | 0.58 [0.45, 0.76] | 1.10 [0.83, 1.47] | NA | NA | 0.94 [0.83, 1.06] | 2.70 [1.72, 4.25] | 0.89 [0.77, 1.03] |
| 95% PI | NA | NA | [0.19, 1.51] | [0.20, 1.74] | [0.31, 3.93] | NA | NA | [0.55, 1.60] | [0.35, 20.75] | [0.42, 1.91] |

TIA: transient ischaemic attack, CVD: cardiovascular disease, CHD: coronary heart disease, ICU: intensive care unit, UK: United Kingdom, MI: myocardial infarction, NR: not reported, CI: confidence interval, PI: prediction interval, NA: not assessed.

*As the models are optimally fit in the development dataset, all OE ratios should be close to 1.

**Figure legends**

Figure 1: Flow chart of study selection.
SR: systematic review, IPD: individual participant data, MA: meta-analysis, NR: not reported, c: concordance, OE: observed expected.

Figure 2: Associations between study characteristics and logit c-statistic with regard to a reference category across 221 external validation studies and 10 different prediction models. Confidence intervals not including 0 are marked with an *. Figure S1 shows these differences on the original scale if we assume a c-statistic of 0.70 in the reference category. For example, for comparability of eligibility criteria, if we assume a c-statistic of 0.70 in the reference category (narrower), this would result in c-statistics of 0.74 [0.72, 0.77], 0.73 [0.66, 0.79], 0.77 [0.68, 0.84], and 0.77 [0.59,0.89] in the categories comparable, mixture, broader, and unclear, respectively.

Figure 3: C-statistic for categories of study design, pooled using univariable meta-regression analyses within each systematic review. N represents the number of external validation studies in a specific category. C diff represents the difference in c-statistic with regard to a reference category (indicated with 'ref').

Figure 4: Associations between study characteristics and ln OE ratio with regard to a reference category across 124 external validation studies and 6 different prediction models. Confidence intervals not including 0 are marked with an *. Figure S4 shows these differences on the original scale if we assume an OE ratio of 1.00 in the reference category. For example, for comparability of eligibility criteria, if we assume an OE ratio of 1.00 in the reference category (narrower), this would result in OE ratios of 0.83 [0.66, 1.05], 1.11 [0.72, 1.70], and 0.92 [0.54, 1.58] in the categories comparable, mixture, and broader, respectively.

Figure 5: Total OE ratio for categories of study design, pooled using univariable meta-regression analyses within each systematic review. N represents the number of external validation studies in a specific category. OE diff represents the difference in OE ratio with regard to a reference category (indicated with 'ref').

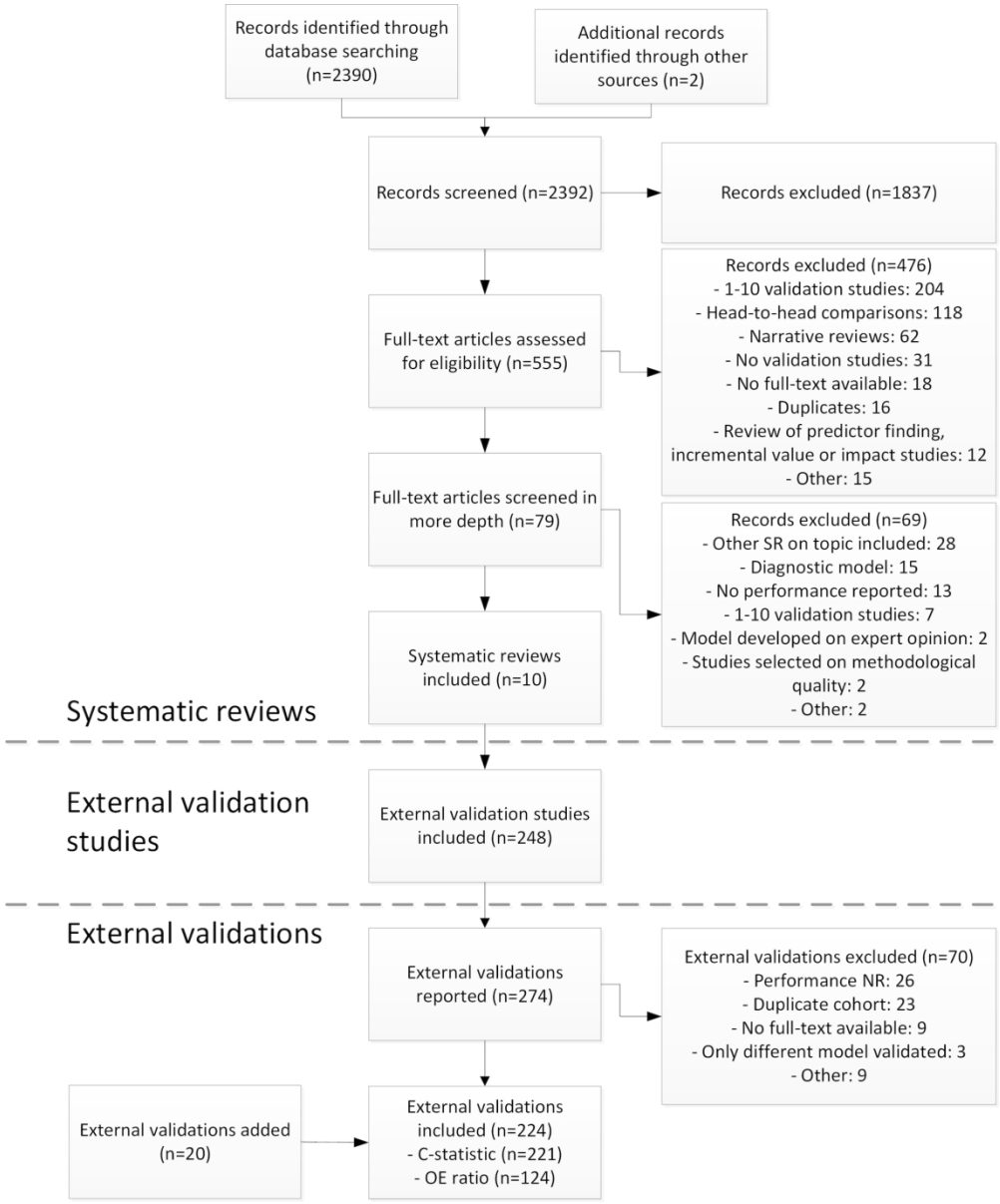



Figure 1: Flow chart of study selection. SR: systematic review, IPD: individual participant data, MA: meta-analysis, NR: not reported, c: concordance, OE: observed expected.

184x222mm (300 x 300 DPI)

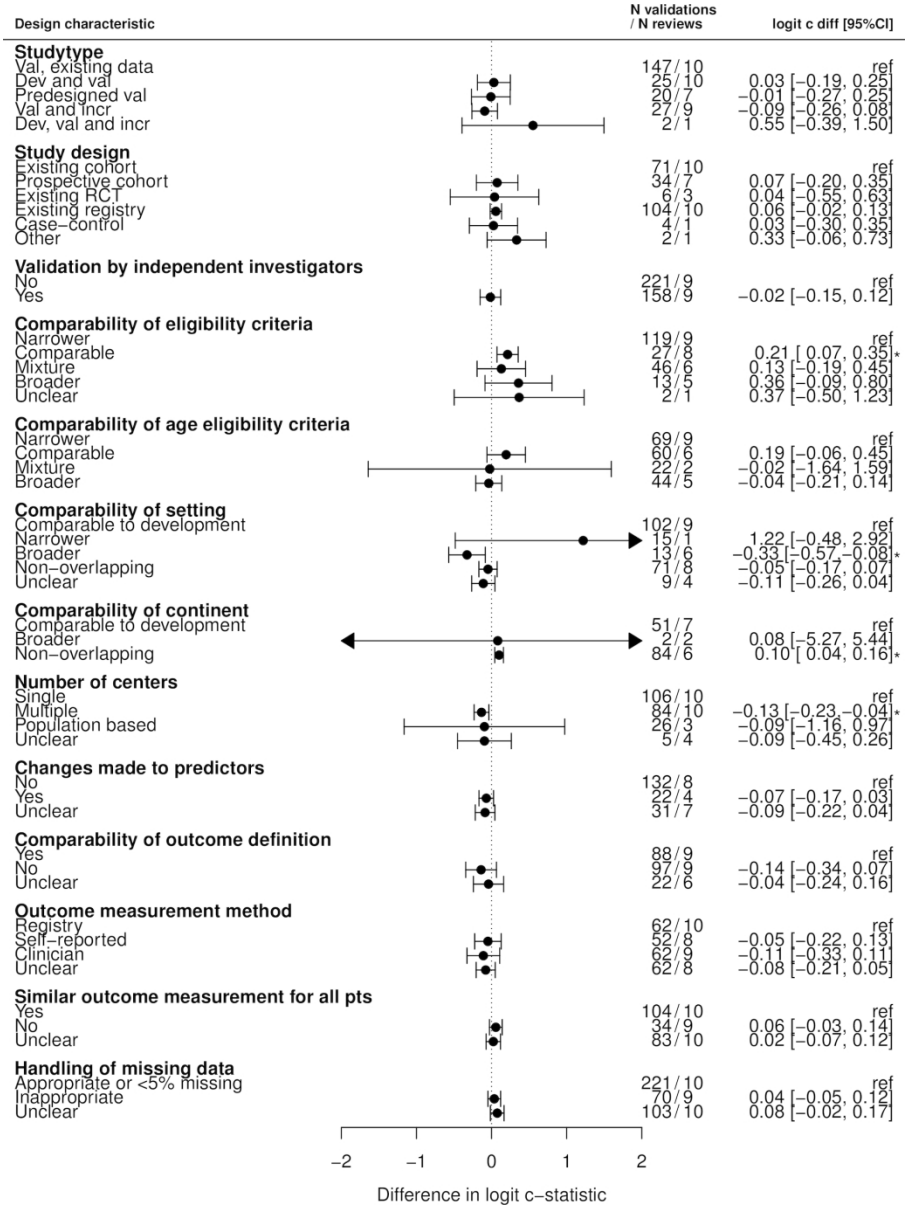



Figure 2: Associations between study characteristics and logit c-statistic with regard to a reference category across 221 external validation studies and 10 different prediction models. Confidence intervals not including 0 are marked with an *. Figure S1 shows these differences on the original scale if we assume a c-statistic of 0.70 in the reference category. For example, for comparability of eligibility criteria, if we assume a c-statistic of 0.70 in the reference category (narrower), this would result in c-statistics of 0.74 [0.72, 0.77], 0.73 [0.66, 0.79], 0.77 [0.68, 0.84], and 0.77 [0.59,0.89] in the categories comparable, mixture, broader, and unclear, respectively.

167x222mm (300 x 300 DPI)

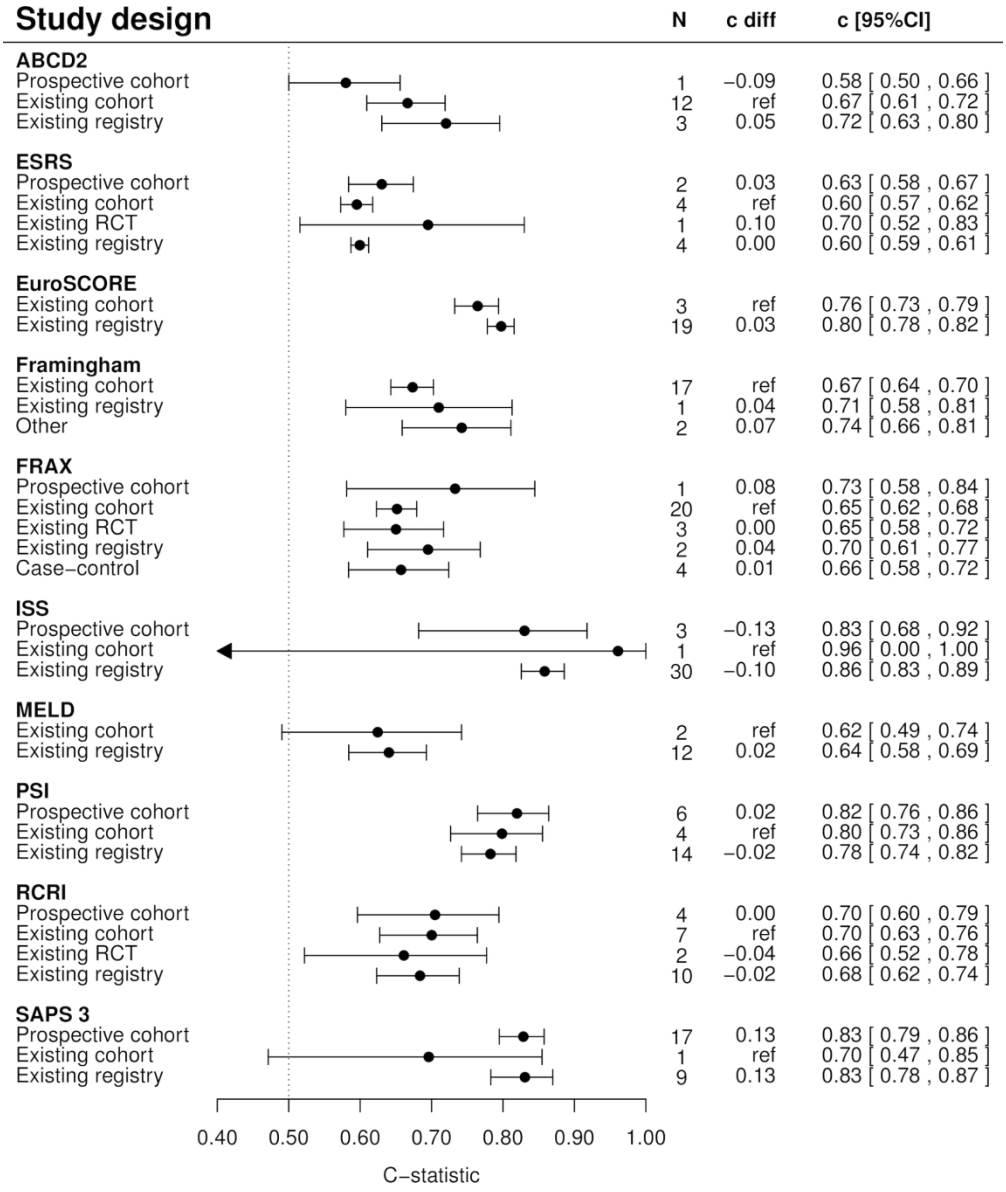



Figure 3: C-statistic for categories of study design, pooled using univariable meta-regression analyses within each systematic review. N represents the number of external validation studies in a specific category. C diff represents the difference in c-statistic with regard to a reference category (indicated with 'ref').

188x222mm (300 x 300 DPI)

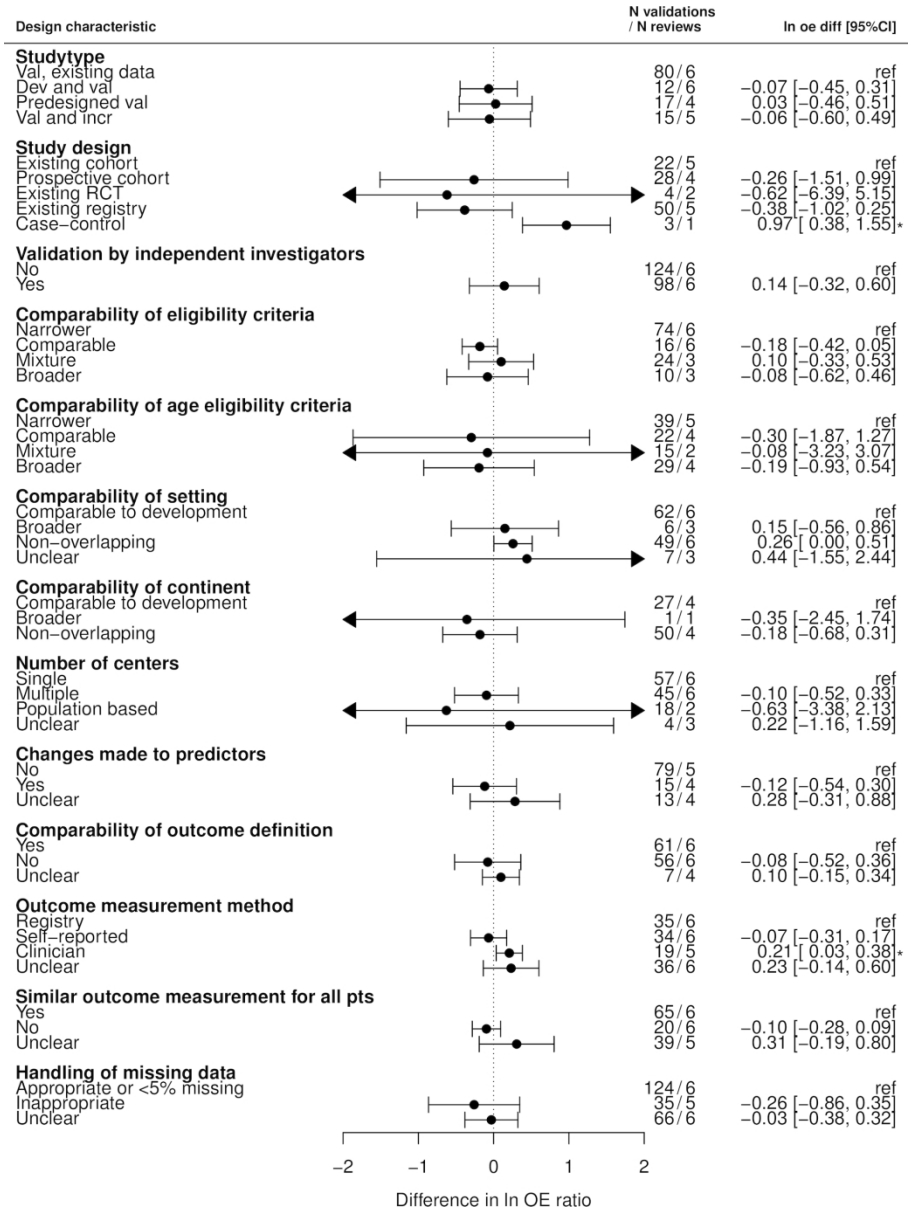



Figure 4: Associations between study characteristics and ln OE ratio with regard to a reference category across 124 external validation studies and 6 different prediction models. Confidence intervals not including 0 are marked with an *. Figure S4 shows these differences on the original scale if we assume an OE ratio of 1.00 in the reference category. For example, for comparability of eligibility criteria, if we assume an OE ratio of 1.00 in the reference category (narrower), this would result in OE ratios of 0.83 [0.66, 1.05], 1.11 [0.72, 1.70], and 0.92 [0.54, 1.58] in the categories comparable, mixture, and broader, respectively.

167x222mm (300 x 300 DPI)

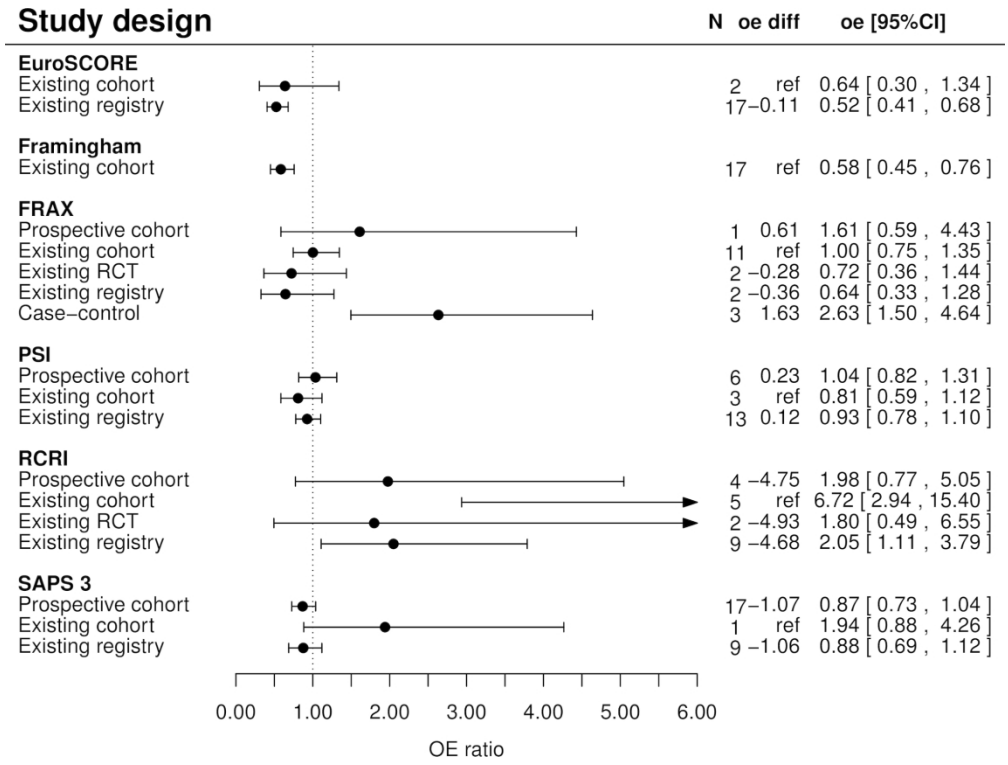



Figure 5: OE ratio for categories of study design, pooled using univariable meta-regression analyses within each systematic review. N represents the number of external validation studies in a specific category. OE diff represents the difference in OE ratio with regard to a reference category (indicated with 'ref').

190x142mm (300 x 300 DPI)

# Supplement to "Empirical evidence on the impact of study characteristics and the performance of prediction models: a meta-epidemiological study"

Johanna A A G Damen, Thomas P A Debray, Romin Pajouheshnia, Johannes B Reitsma, Rob J P M Scholten, Karel G M Moons, Lotty Hooft

**Content**

**Supplement 1: Search string and selection criteria**

The systematic reviews included in this article have been selected from a database with reviews of prediction models. To construct this database, the following search string was used to search Pubmed (last update: October 16th, 2018):

(clinical prediction[ti] OR
risk calculator*[ti] OR
risk index[ti] OR
risk indices[ti] OR
risk model*[ti] OR
risk prediction[ti] OR
risk score*[ti] OR
risk stratification[ti] OR
predictive model*[ti] OR
prediction model*[ti] OR
prediction rule*[tiab] OR
prognostic index[ti] OR
prognostic indices[ti] OR
prognostic model*[ti] OR
scoring system*[ti]) AND
(review[Publication Type] OR
review[ti] OR
critical appraisal[ti] OR
Bibliography[Publication Type] OR
Meta-analysis[Publication Type]) NOT
(Editorial[Publication Type] OR
Letter[Publication Type] OR
News[Publication Type])

All references identified with this search string were screened for eligibility to the database by one reviewer (Gary Collins) based on title, abstract and, if necessary, full text. The following in- and exclusion criteria were applied:

Inclusion criteria
- Review (narrative or systematic) of prediction models
- Both diagnostic and prognostic prediction models
- Validation studies in which multiple prediction models have been validated (head-to-head comparison)
- Review of impact studies of prediction models
- Protocol of systematic review of prediction models

Exclusion criteria
- Comments, editorials

**Supplement 2: Description of items extracted from studies and included in analyses**

| Item | Extracted from studies | Categorization / handling in analyses | Description / examples |
|---|---|---|---|
| Validated model | ABCD2, ESRS, EuroSCORE, Framingham Wilson, FRAX, ISS, MELD, PSI, RCRI, SAPS 3 | - | - |
| Study type | Predesigned validation study | Predesigned validation study | Study designed with the aim of validating a prediction model |
| | Validation study using existing data | Validation study using existing data | Study in which a prediction model is validated using a dataset collected for a different purpose than validating the model |
| | Development of new model and validation of different model | Development of new model and validation of different model | Study in which a model is developed and a model is validated |
| | Validation and incremental value | Validation and incremental value | Study in which a model is validated and in which the added value of one or more predictors is assessed |
| | Development, validation, and incremental value study | Development, validation, and incremental value study | Combination of the two above |
| Independent investigators | Yes | Yes | None of the authors of the development study was listed as author in the external validation study |
| | No | No | One or more of the authors of the development study was listed as author in the external validation study |
| Study design | Prospective cohort | Prospective cohort | |
| | Existing cohort | Existing cohort | |
| | Existing RCT | Existing RCT | |
| | Existing registry / medical records | Existing registry | |
| | Case-control | Case-control | |
| | Other (specify) | Other | |
| Eligibility criteria for participants | Copy/paste eligibility criteria of validation study | Comparable | Eligibility criteria comparable to development study |

| Item | Extracted from studies | Categorization / handling in analyses | Description / examples |
|---|---|---|---|
| | | Narrower | People included in the development study excluded in the validation study |
| | | Broader | People excluded in the development stud included in the validation study |
| | | Mixture | Combination of narrower and broader |
| | | Unclear | |
| Setting | Primary care<br>Secondary care<br>Tertiary care<br>Population based<br>Screening<br>Mixed<br>Unclear | Comparable | Same setting as development study |
| | | Broader | Same setting as development study, and participants from additional settings recruited |
| | | Non-overlapping | Setting in development study differs from validation study |
| | | Unclear | |
| Study dates | Start year of recruitment<br>End year of recruitment | Continuous, standardized per systematic review | |
| Prediction horizon | Time period for which predictions were made, eg, 10 years. | Continuous, standardized per systematic review | |
| Geographical location | Country and continent | Comparable | Model validated in the same continent as the development study |
| | | Broader | Model validated in the same and additional continents as the development study |
| | | Non-overlapping | Model validated in a different continent than the development study |
| Number of centres | Number of centres (numerical) | Single | |
| | | Multiple | |
| | | Population based | Participants not recruited at medical centres, but, for example, from a specific geographic area (eg, all individuals living in Framingham, US) |
| | | Unclear | |
| Case-mix: age mean | Mean and SD of age of | Continuous, standardized per | |

| Item | Extracted from studies | Categorization / handling in analyses | Description / examples |
|---|---|---|---|
| and sd | participants included in the study, or other available information about age distribution | systematic review | |
| Case-mix: gender | Percentage of men included in a study | Continuous, standardized per systematic review | |
| Predictors | Were predictors deleted from the model, or were predictors substituted with different predictors. | Yes | Changes made to predictors |
| | | No | No changes made to predictors |
| | | Unclear | |
| Predicted outcome | Full definition, including ICD-codes | Comparable | Outcome definition comparable to development study |
| | | Not comparable | Outcome definition not comparable to development study |
| | | Unclear | |
| Outcome - measurement method | Measurement method (eg, self-reported, interviews, expert panel), differences in outcome measurement between participants in the study | Yes | Outcome measurement similar for all participant |
| | | No | Systematic differences in outcome measurement between participants |
| | | Unclear | |
| Missing data | Number of participants with missing data, method of handling missing data | Appropriate | Missing data handled using multiple imputation, or <5% missing data (arbitrary cut-off) |
| | | Inappropriate | Missing data not handled using multiple imputation (eg, complete-case analysis, mean imputation), and >=5% missing data |
| | | Unclear | Unclear handling of missing data, and >=5% missing data |
| Number of participants | | Continuous, standardized per systematic review | |
| Number of events | | Continuous, standardized per systematic review | |

| Item | Extracted from studies | Categorization / handling in analyses | Description / examples |
|---|---|---|---|
| Model updating | Was the model altered before validating it, eg, using intercept recalibration. | NA | |
| Performance - c-statistic | C-statistic, AUC, 95% confidence intervals or SE | Logit transformation[1] | |
| Performance - total OE ratio | OE ratio, predicted risks, presence of calibration plots or tables, 95% confidence intervals or SE | Ln transformation[1] | |

SD: standard deviation, NA: not applicable, C-statistic: concordance statistic, AUC: area under the receiver operating curve, SE: standard error, OE ratio: observed expected ratio.

Information regarding c-statistics and total OE ratios when not reported was sometimes restored from other information reported in the paper. If the precision of the c-statistic was not reported, we estimated this from the c-statistic and sample size of the study, using the formula described by Newcombe and Hanley.[2 3] Various equations were used to estimate the standard error of the OE ratio, depending on which information was reported. All equations (as numbered) are described in the appendix of Debray et al.[4] If the SE of the OE ratio was reported, we used equation 16 to estimate the SE of ln(OE), if the observed event risk (Po), the expected event risk (Pe), and the SE of Po were reported, we used equation 51, and if only Po and Pe were reported we used equation 27.

**Supplement 3: Statistical analyses**

First we pooled the total OE ratio and c-statistic within each systematic review. Based on previous recommendations,[1 4] we pooled the log OE ratio and logit c-statistic using random-effects meta-analysis accounting for the presence of between-study heterogeneity, weighted by the inverse of the variance. We calculated 95% confidence intervals (CI) and (approximate) 95% prediction intervals (PI) to quantify uncertainty and the presence of between-study heterogeneity. The Hartung-Knapp-Sidik-Jonkman (HKSJ) method was used when calculating 95% CIs.[5] The 95% PI was calculated using the equation described previously.[4] The CI indicates the precision of the summary performance estimate and the PI provides boundaries on the likely performance in future model validation studies that are comparable to the studies included in the meta-analysis, and can thus be seen as an indication of model generalizability.[6]

To study the possible association between study characteristics and predictive performance, we used a two-stepped approach. In the first stage, we fitted a univariable meta-regression model (ie, a separate model for every study characteristic) within every systematic review, with the logit c-statistic or log OE ratio as outcome variable. This model was fitted with intercept term. Therefore, the effect estimates obtained from this meta-regression model indicate the difference in logit c-statistic or log OE ratio between a certain category of a study characteristic and a chosen reference category of that characteristic. As a reference category, we chose the category that was present in the highest number of systematic reviews allowing the inclusion of as many data as possible.

In the second stage, these effect estimates were pooled with a random effects meta-analysis model. This reflected the influence of the study characteristic on model performance over all systematic reviews. For continuous study characteristics, the intercept term and beta-coefficient from the first stage were jointly pooled across reviews using bivariate meta-analysis.[4 6] For categorical study characteristics the data available were not sufficient for the complexity of a multivariate model, so every category was pooled in a separate (univariate) meta-analysis.

As the estimates obtained with this approach are on the transformed scale (ie, the difference in logit c-statistic or log OE ratio between one category and the reference category), we transformed these back assuming a c-statistic of 0.7 or an OE ratio of 1.00 in the reference category. Also, we performed a second analysis where we again fitted a univariable meta-regression model, with the logit c-statistic or log OE ratio as outcome variable, but now without intercept term. This analysis enables the calculation

of an effect estimate for every category of a study characteristic and to back transform this to the original scale, yielding a pooled c-statistic or pooled OE ratio for each category of a study characteristic. We planned to perform multivariable analyses to assess the association between various study characteristics in combination and the performance of prediction models, but due to the paucity of data we were not able to do so.

All analyses were performed in R version 3.3.2,[7] using the packages metafor,[8] mvmeta,[9] metamisc,[10] and lme4.[11]

Table S1: Description of study characteristics and quality of reporting within each systematic review

Categorical variables

| | ABCD2 | ESRS | EuroSCORE | Framingham | FRAX | ISS | MELD | PSI | RCRI | SAPS 3 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Studytype** | | | | | | | | | | |
| Validation study using existing data | 9 (56%) | 7 (64%) | 21 (95%) | 18 (78%) | 26 (87%) | 24 (71%) | 10 (71%) | 16 (67%) | 11 (48%) | 8 (30%) |
| Development of new model and validation of different model | 2 (12%) | 2 (18%) | 1 (5%) | 2 (9%) | 1 (3%) | 4 (12%) | 3 (21%) | 2 (8%) | 2 (9%) | 6 (22%) |
| Development, validation, and incremental value study | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 2 (6%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Predesigned validation study | 1 (6%) | 1 (9%) | 0 (0%) | 0 (0%) | 1 (3%) | 1 (3%) | 0 (0%) | 5 (21%) | 1 (4%) | 10 (37%) |
| Validation and incremental value | 4 (25%) | 1 (9%) | 0 (0%) | 3 (13%) | 2 (7%) | 3 (9%) | 1 (7%) | 1 (4%) | 9 (39%) | 3 (11%) |
| **Study design** | | | | | | | | | | |
| Existing cohort | 12 (75%) | 4 (36%) | 3 (14%) | 20 (87%) | 20 (67%) | 1 (3%) | 2 (14%) | 4 (17%) | 7 (30%) | 1 (4%) |
| Prospective cohort | 1 (6%) | 2 (18%) | 0 (0%) | 0 (0%) | 1 (3%) | 3 (9%) | 0 (0%) | 6 (25%) | 4 (17%) | 17 (63%) |
| Existing RCT | 0 (0%) | 1 (9%) | 0 (0%) | 0 (0%) | 3 (10%) | 0 (0%) | 0 (0%) | 0 (0%) | 2 (9%) | 0 (0%) |
| Existing registry | 3 (19%) | 4 (36%) | 19 (86%) | 1 (4%) | 2 (7%) | 30 (88%) | 12 (86%) | 14 (58%) | 10 (43%) | 9 (33%) |
| Case-control | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 4 (13%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Other | 0 (0%) | 0 (0%) | 0 (0%) | 2 (9%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| **Validation by independent investigators** | | | | | | | | | | |
| No | 6 (38%) | 3 (27%) | 3 (14%) | 10 (43%) | 17 (57%) | 2 (6%) | 0 (0%) | 7 (29%) | 1 (4%) | 2 (7%) |
| Yes | 10 (62%) | 8 (73%) | 19 (86%) | 13 (57%) | 13 (43%) | 32 (94%) | 14 (100%) | 17 (71%) | 22 (96%) | 25 (93%) |
| **Comparability of eligibility criteria** | | | | | | | | | | |
| Narrower | 6 (38%) | 2 (18%) | 18 (82%) | 18 (78%) | 28 (93%) | 22 (65%) | 0 (0%) | 4 (17%) | 4 (17%) | 20 (74%) |
| Comparable | 4 (25%) | 0 (0%) | 2 (9%) | 3 (13%) | 2 (7%) | 5 (15%) | 0 (0%) | 1 (4%) | 3 (13%) | 7 (26%) |
| Mixture | 5 (31%) | 9 (82%) | 0 (0%) | 2 (9%) | 0 (0%) | 3 (9%) | 0 (0%) | 16 (67%) | 11 (48%) | 0 (0%) |
| Broader | 1 (6%) | 0 (0%) | 2 (9%) | 0 (0%) | 0 (0%) | 2 (6%) | 0 (0%) | 3 (12%) | 5 (22%) | 0 (0%) |
| Non-overlapping | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 14 (100%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Unclear | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 2 (6%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| **Comparability of age eligibility criteria** | | | | | | | | | | |
| Narrower | 1 (6%) | 2 (18%) | 0 (0%) | 17 (74%) | 9 (30%) | 13 (38%) | 10 (71%) | 2 (8%) | 1 (4%) | 17 (63%) |
| Comparable | 15 (94%) | 0 (0%) | 22 (100%) | 0 (0%) | 1 (3%) | 21 (62%) | 4 (29%) | 15 (62%) | 4 (17%) | 4 (15%) |
| Mixture | 0 (0%) | 0 (0%) | 0 (0%) | 6 (26%) | 16 (53%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Broader | 0 (0%) | 9 (82%) | 0 (0%) | 0 (0%) | 4 (13%) | 0 (0%) | 0 (0%) | 7 (29%) | 18 (78%) | 6 (22%) |
| **Setting** | | | | | | | | | | |

Damen et al. Empirical evidence on the impact of study characteristics and the performance of prediction models: a meta-epidemiological study

| | ABCD2 | ESRS | EuroSCORE | Framingham | FRAX | ISS | MELD | PSI | RCRI | SAPS 3 |
|---|---|---|---|---|---|---|---|---|---|---|
| Primary care | 3 (19%) | 0 (0%) | 0 (0%) | 7 (30%) | 3 (10%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Secondary care | 12 (75%) | 9 (82%) | 16 (84%) | 0 (0%) | 5 (17%) | 18 (82%) | 6 (75%) | 18 (90%) | 12 (86%) | 4 (40%) |
| Tertiary care | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Population based | 0 (0%) | 1 (9%) | 0 (0%) | 15 (65%) | 17 (59%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Screening | 0 (0%) | 0 (0%) | 0 (0%) | 1 (4%) | 1 (3%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Mixed | 1 (6%) | 1 (9%) | 2 (11%) | 0 (0%) | 0 (0%) | 4 (18%) | 1 (12%) | 2 (10%) | 2 (14%) | 2 (20%) |
| Unclear | 0 (0%) | 0 (0%) | 1 (5%) | 0 (0%) | 3 (10%) | 0 (0%) | 1 (12%) | 0 (0%) | 0 (0%) | 4 (40%) |
| **Comparability of setting** | | | | | | | | | | |
| Comparable | 1 (6%) | 0 (0%) | 16 (73%) | 15 (65%) | 17 (57%) | 18 (53%) | 6 (43%) | 18 (75%) | 9 (39%) | 4 (15%) |
| Narrower | 15 (94%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Broader | 0 (0%) | 0 (0%) | 2 (9%) | 0 (0%) | 0 (0%) | 4 (12%) | 1 (7%) | 2 (8%) | 2 (9%) | 2 (7%) |
| Non-overlapping | 0 (0%) | 11 (100%) | 3 (14%) | 8 (35%) | 10 (33%) | 12 (35%) | 6 (43%) | 4 (17%) | 12 (52%) | 17 (63%) |
| Unclear | 0 (0%) | 0 (0%) | 1 (5%) | 0 (0%) | 3 (10%) | 0 (0%) | 1 (7%) | 0 (0%) | 0 (0%) | 4 (15%) |
| **Continent** | | | | | | | | | | |
| Africa | 0 (0%) | 0 (0%) | 1 (5%) | 0 (0%) | 0 (0%) | 1 (3%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Asia | 0 (0%) | 2 (18%) | 4 (18%) | 2 (9%) | 4 (13%) | 7 (21%) | 1 (7%) | 3 (12%) | 2 (9%) | 5 (19%) |
| Australia | 0 (0%) | 0 (0%) | 1 (5%) | 0 (0%) | 5 (17%) | 2 (6%) | 0 (0%) | 3 (12%) | 1 (4%) | 1 (4%) |
| Europe | 8 (50%) | 7 (64%) | 10 (45%) | 10 (43%) | 9 (30%) | 7 (21%) | 7 (50%) | 11 (46%) | 13 (57%) | 9 (33%) |
| North America | 8 (50%) | 1 (9%) | 5 (23%) | 11 (48%) | 11 (37%) | 17 (50%) | 3 (21%) | 6 (25%) | 6 (26%) | 3 (11%) |
| South America | 0 (0%) | 0 (0%) | 1 (5%) | 0 (0%) | 0 (0%) | 0 (0%) | 3 (21%) | 0 (0%) | 0 (0%) | 9 (33%) |
| Combination | 0 (0%) | 1 (9%) | 0 (0%) | 0 (0%) | 1 (3%) | 0 (0%) | 0 (0%) | 1 (4%) | 1 (4%) | 0 (0%) |
| **Comparability of continent** | | | | | | | | | | |
| Comparable | 0 (0%) | 0 (0%) | 10 (45%) | 11 (48%) | 0 (0%) | 17 (50%) | 3 (21%) | 6 (25%) | 6 (26%) | 0 (0%) |
| Narrower | 16 (100%) | 8 (73%) | 0 (0%) | 0 (0%) | 30 (100%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 27 (100%) |
| Broader | 0 (0%) | 1 (9%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (4%) | 1 (4%) | 0 (0%) |
| Non-overlapping | 0 (0%) | 2 (18%) | 12 (55%) | 12 (52%) | 0 (0%) | 17 (50%) | 11 (79%) | 17 (71%) | 16 (70%) | 0 (0%) |
| **Number of centres** | | | | | | | | | | |
| Single | 9 (56%) | 4 (36%) | 12 (55%) | 3 (13%) | 5 (17%) | 18 (53%) | 12 (86%) | 14 (58%) | 17 (74%) | 13 (48%) |
| Multiple | 6 (38%) | 7 (64%) | 9 (41%) | 6 (26%) | 9 (30%) | 15 (44%) | 2 (14%) | 10 (42%) | 6 (26%) | 14 (52%) |
| Population based | 1 (6%) | 0 (0%) | 0 (0%) | 12 (52%) | 15 (50%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Unclear | 0 (0%) | 0 (0%) | 1 (5%) | 2 (9%) | 1 (3%) | 1 (3%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |

Damen et al. Empirical evidence on the impact of study characteristics and the performance of prediction models: a meta-epidemiological study

| | ABCD2 | ESRS | EuroSCORE | Framingham | FRAX | ISS | MELD | PSI | RCRI | SAPS 3 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Changes made to predictors** | | | | | | | | | | |
| No | 16 (100%) | 10 (91%) | 13 (59%) | 23 (100%) | 20 (67%) | 21 (62%) | 12 (86%) | 13 (54%) | 17 (74%) | 26 (96%) |
| Yes | 0 (0%) | 0 (0%) | 5 (23%) | 0 (0%) | 10 (33%) | 0 (0%) | 0 (0%) | 5 (21%) | 2 (9%) | 0 (0%) |
| Unclear | 0 (0%) | 1 (9%) | 4 (18%) | 0 (0%) | 0 (0%) | 13 (38%) | 2 (14%) | 6 (25%) | 4 (17%) | 1 (4%) |
| **Comparability of outcome definition** | | | | | | | | | | |
| No | 8 (50%) | 3 (27%) | 9 (41%) | 13 (57%) | 16 (53%) | 7 (21%) | 14 (100%) | 5 (21%) | 5 (22%) | 24 (89%) |
| Yes | 3 (19%) | 8 (73%) | 11 (50%) | 4 (17%) | 13 (43%) | 19 (56%) | 0 (0%) | 18 (75%) | 18 (78%) | 3 (11%) |
| Unclear | 5 (31%) | 0 (0%) | 2 (9%) | 6 (26%) | 1 (3%) | 8 (24%) | 0 (0%) | 1 (4%) | 0 (0%) | 0 (0%) |
| **Outcome measurement method** | | | | | | | | | | |
| Self-reported | 3 (19%) | 8 (73%) | 2 (9%) | 15 (65%) | 18 (60%) | 0 (0%) | 1 (7%) | 6 (25%) | 1 (4%) | 2 (7%) |
| Clinician | 6 (38%) | 1 (9%) | 2 (9%) | 2 (9%) | 1 (3%) | 4 (12%) | 0 (0%) | 2 (8%) | 10 (43%) | 6 (22%) |
| Registry | 3 (19%) | 2 (18%) | 5 (23%) | 4 (17%) | 8 (27%) | 13 (38%) | 3 (21%) | 9 (38%) | 6 (26%) | 9 (33%) |
| Unclear | 4 (25%) | 0 (0%) | 13 (59%) | 2 (9%) | 3 (10%) | 17 (50%) | 10 (71%) | 7 (29%) | 6 (26%) | 10 (37%) |
| **Similar outcome measurement for all patients** | | | | | | | | | | |
| Yes | 9 (56%) | 3 (27%) | 11 (50%) | 6 (26%) | 14 (47%) | 18 (53%) | 2 (14%) | 12 (50%) | 13 (57%) | 16 (59%) |
| No | 1 (6%) | 5 (45%) | 1 (5%) | 16 (70%) | 3 (10%) | 3 (9%) | 0 (0%) | 3 (12%) | 4 (17%) | 1 (4%) |
| Unclear | 6 (38%) | 3 (27%) | 10 (45%) | 1 (4%) | 13 (43%) | 13 (38%) | 12 (86%) | 9 (38%) | 6 (26%) | 10 (37%) |
| **Method for handling of missing data** | | | | | | | | | | |
| Complete case analysis | 4 (25%) | 8 (73%) | 3 (14%) | 11 (48%) | 19 (63%) | 18 (53%) | 7 (50%) | 1 (4%) | 5 (22%) | 8 (30%) |
| Mean/median imputation | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (4%) |
| Multiple imputation | 1 (6%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| NA | 3 (19%) | 1 (9%) | 1 (5%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (4%) | 1 (4%) | 2 (7%) |
| Other | 0 (0%) | 0 (0%) | 1 (5%) | 0 (0%) | 2 (7%) | 0 (0%) | 1 (7%) | 8 (33%) | 0 (0%) | 7 (26%) |
| Unclear | 8 (50%) | 2 (18%) | 17 (77%) | 12 (52%) | 9 (30%) | 16 (47%) | 6 (43%) | 14 (58%) | 17 (74%) | 9 (33%) |
| **Handling of missing data** | | | | | | | | | | |
| Appropriate or <5% missing | 8 (50%) | 4 (36%) | 2 (9%) | 3 (13%) | 4 (13%) | 6 (18%) | 4 (29%) | 4 (17%) | 7 (30%) | 6 (22%) |
| Inappropriate | 1 (6%) | 5 (45%) | 3 (14%) | 8 (35%) | 17 (57%) | 13 (38%) | 4 (29%) | 8 (33%) | 0 (0%) | 12 (44%) |
| Unclear | 7 (44%) | 2 (18%) | 17 (77%) | 12 (52%) | 9 (30%) | 15 (44%) | 6 (43%) | 12 (50%) | 16 (70%) | 9 (33%) |

Continuous variables

| | **ABCD2** | **ESRS** | **EuroSCORE** | **Framingham** | **FRAX** | **ISS** | **MELD** | **PSI** | **RCRI** | **SAPSIII** |
|---|---|---|---|---|---|---|---|---|---|---|
| Year start recruitment | 2002 (2000-2003) NR=0 | 2007 (2004-2007) NR=0 | 1998 (1995-2001) NR=1 | 1989 (1983-1994) NR=0 | 1994 (1990-1998) NR=3 | 1996 (1993-1998) NR=1 | 2000 (1998-2004) NR=0 | 2000 (1998-2002) NR=0 | 2000 (1994-2002) NR=4 | 2006 (2006-2007) NR=0 |
| Year end recruitment | 2005 (2003-2007) NR=0 | 2008 (2006-2008) NR=0 | 2002 (1999-2005) NR=1 | 1993 (1988-1998) NR=0 | 1997 (1993-2006) NR=8 | 2000 (1996-2003) NR=2 | 2006 (2004-2007) NR=0 | 2002 (2000-2003) NR=0 | 2002 (2000-2005) NR=4 | 2007 (2006-2009) NR=0 |
| Percentage missings | 0.95 (0.00-5.00) NR=7 | 5.12 (1.99-17.80) NR=2 | 6.40 (1.50-11.83) NR=18 | 4.90 (2.70-9.80) NR=18 | 30.25 (2.75-33.80) NR=16 | 9.05 (2.40-14.65) NR=20 | 4.05 (2.73-10.93) NR=8 | 0.52 (0.07-9.26) NR=18 | 1.00 (0.09-1.91) NR=16 | 5.85 (0.52-18.93) NR=15 |
| Number of participants | 304 (204-691) NR=0 | 1257 (712-2594) NR=0 | 1730 (873-4518) NR=2 | 2399 (928-4609) NR=0 | 2210 (889-6586) NR=0 | 2590 (960-20713) NR=0 | 418 (118-483) NR=0 | 730 (326-970) NR=1 | 496 (180-1480) NR=0 | 864 (485-1856) NR=0 |
| Number of events | 9 (3-18) NR=0 | 92 (60-134) NR=0 | 36 (13-87) NR=2 | 92 (72-160) NR=1 | 250 (86-581) NR=0 | 256 (113-1660) NR=2 | 49 (22-112) NR=0 | 54 (28-111) NR=1 | 31 (14-76) NR=0 | 180 (124-311) NR=1 |
| Age mean | 67.4 (64.1-70.0) NR=5 | 68.3 (67.1-71.5) NR=3 | 63.9 (62.5-65.2) NR=2 | 54.6 (50.9-58.3) NR=2 | 66.8 (63.0-71.3) NR=1 | 38.1 (32.4-41.3) NR=10 | 51.8 (49.1-53.0) NR=0 | 66.2 (64.0-69.3) NR=2 | 67.8 (66.0-71.9) NR=2 | 62.2 (60.8-64.8) NR=1 |
| Age sd | 13.8 (13.0-14.9) NR=5 | 12.4 (12.0-13.0) NR=1 | 9.3 (9.0-10.6) NR=8 | 7.3 (4.1-9.4) NR=0 | 8.3 (5.9-9.8) NR=0 | 20.9 (18.1-24.8) NR=2 | 10.0 (9.6-12.0) NR=1 | 17.8 (17.0-20.1) NR=3 | 10.0 (8.8-12.5) NR=4 | 17.0 (15.4-19.0) NR=3 |
| Gender percentage men | 47 (45-53) NR=2 | 57 (55-59) NR=1 | 77 (71-79) NR=1 | 100 (100-100) NR=0 | 0 (0-0) NR=0 | 71 (64-75) NR=11 | 68 (63-69) NR=0 | 57 (53-64) NR=1 | 67 (52-76) NR=0 | 59 (55-64) NR=0 |

Values represent median (IQR), number of missing values.

Figure S1: Associations between categorical variables and c-statistic

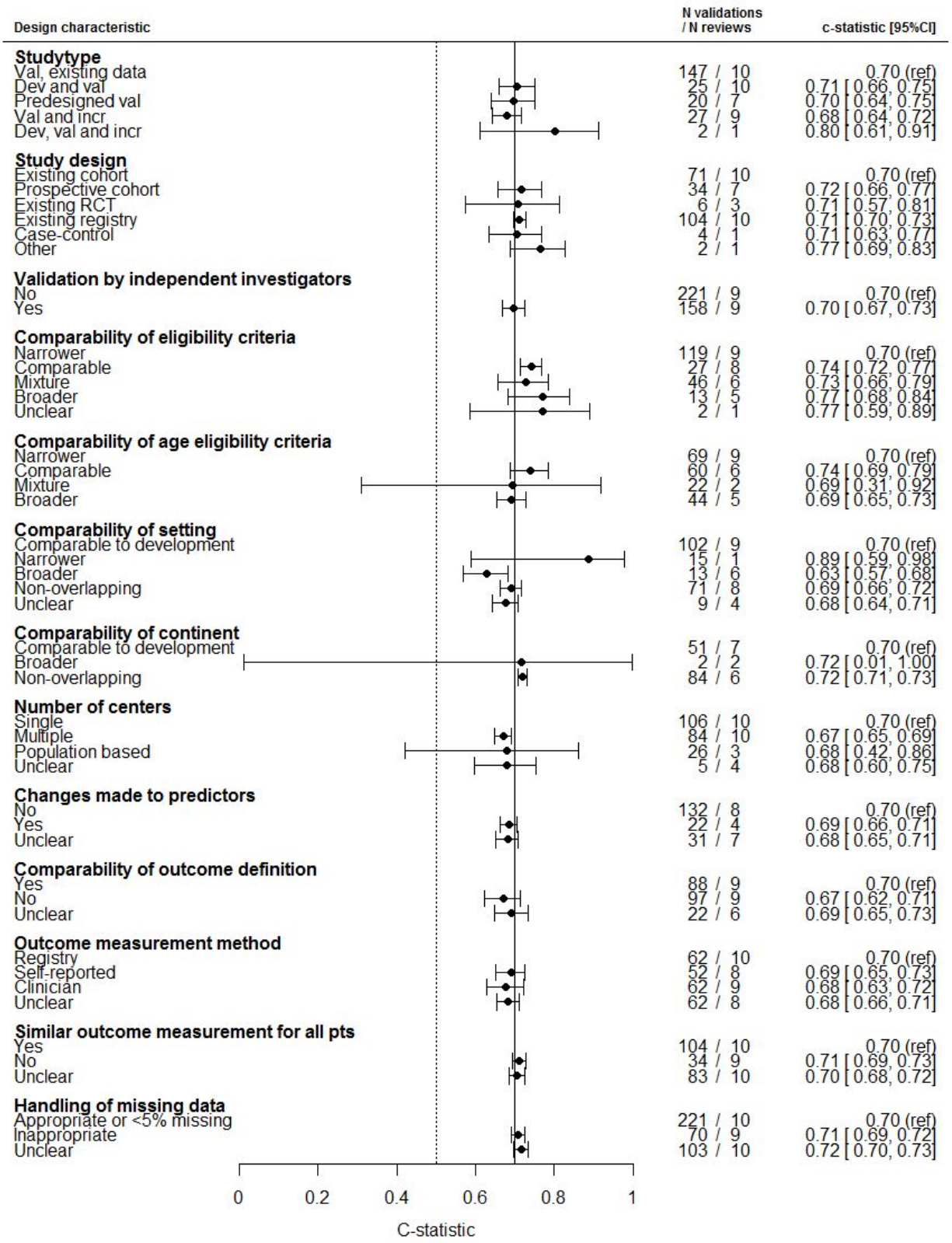
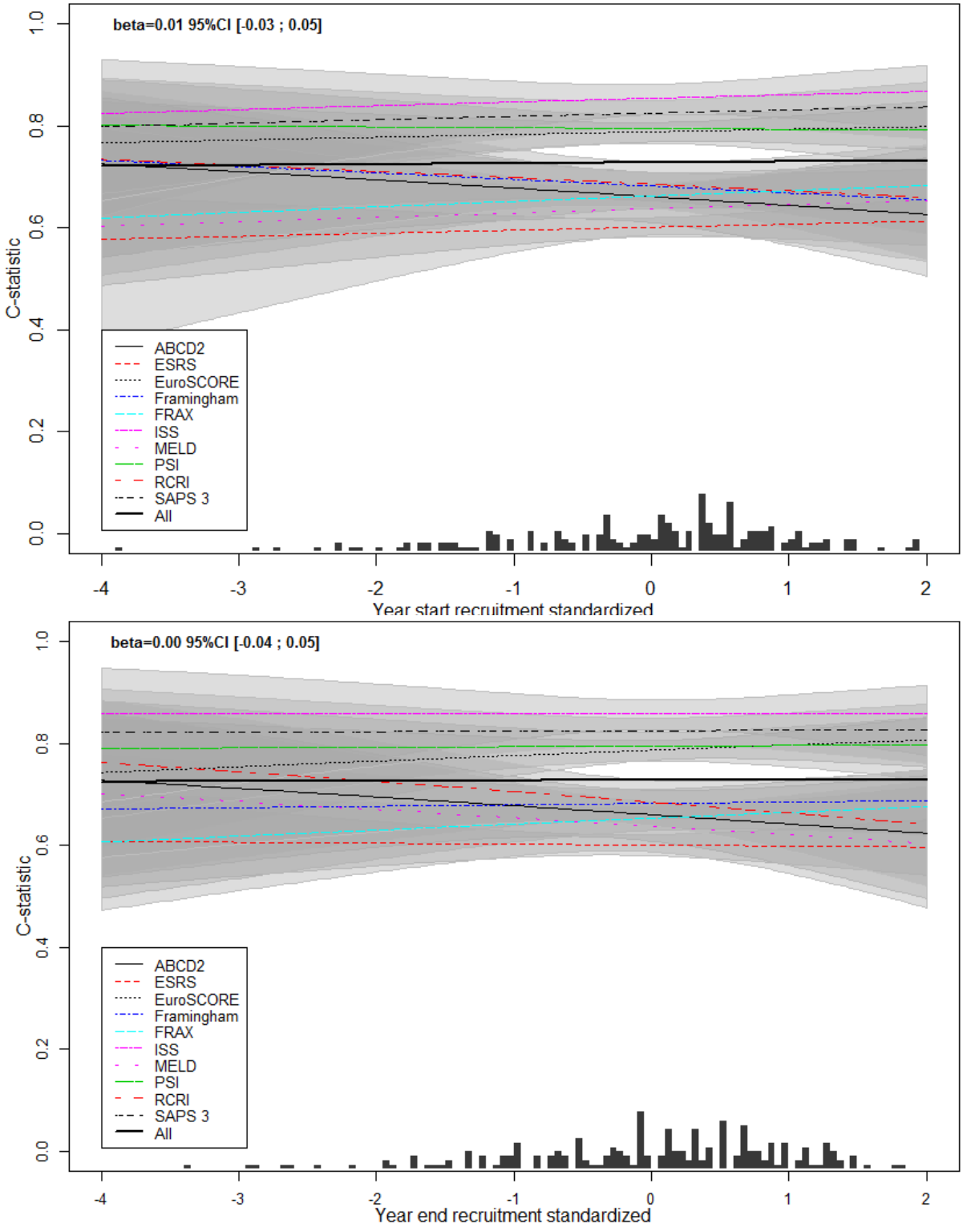
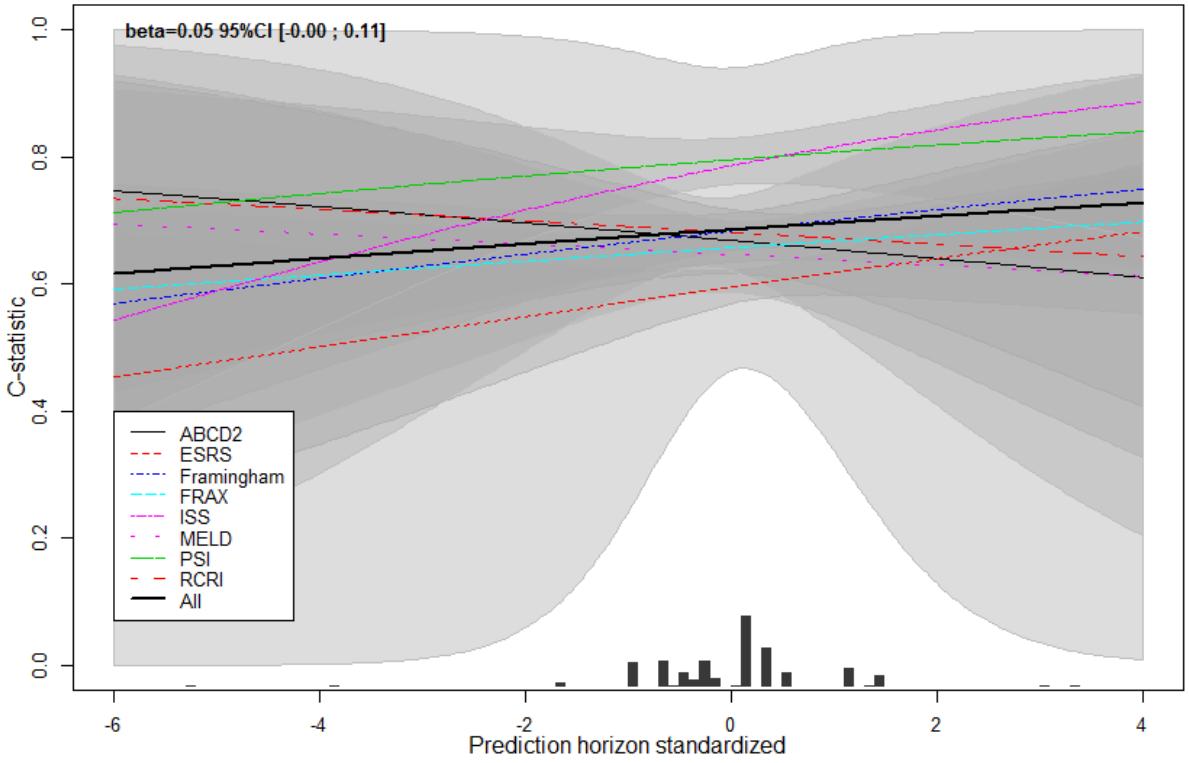
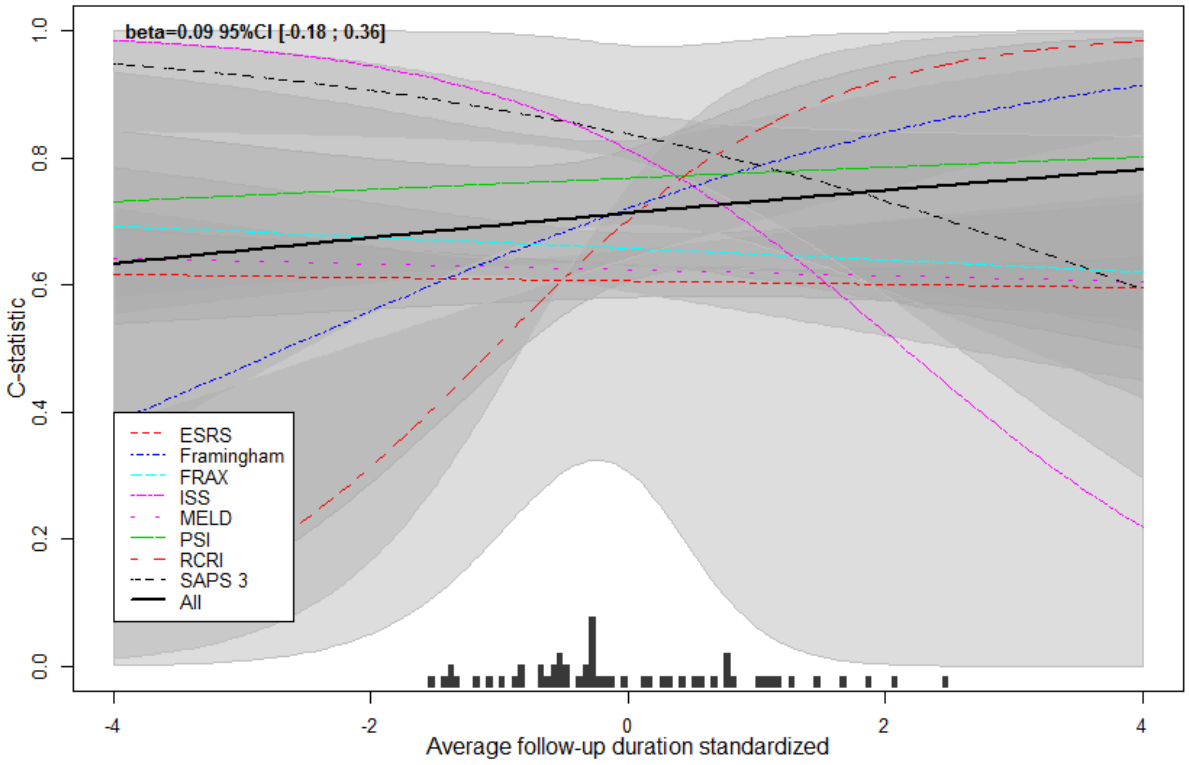


Figure S2: Associations between continuous variables and c-statistic

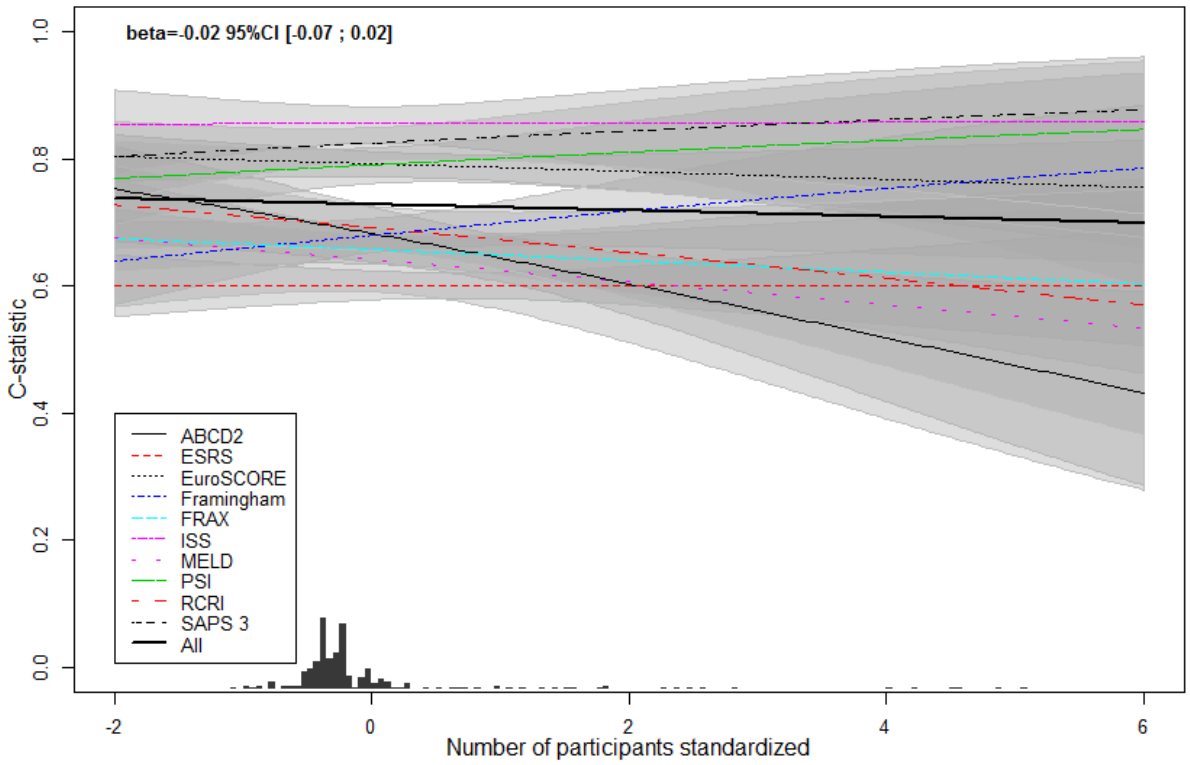


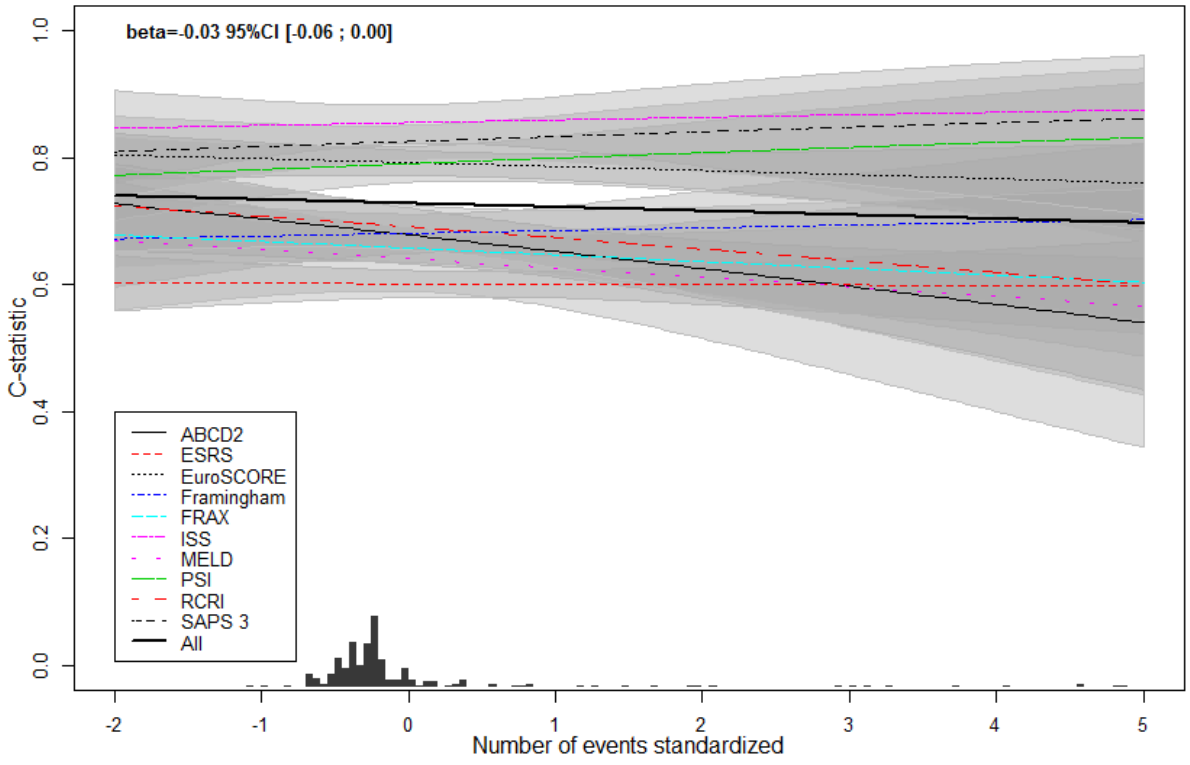
Damen et al. Empirical evidence on the impact of study characteristics and the performance of prediction models: a meta-epidemiological study



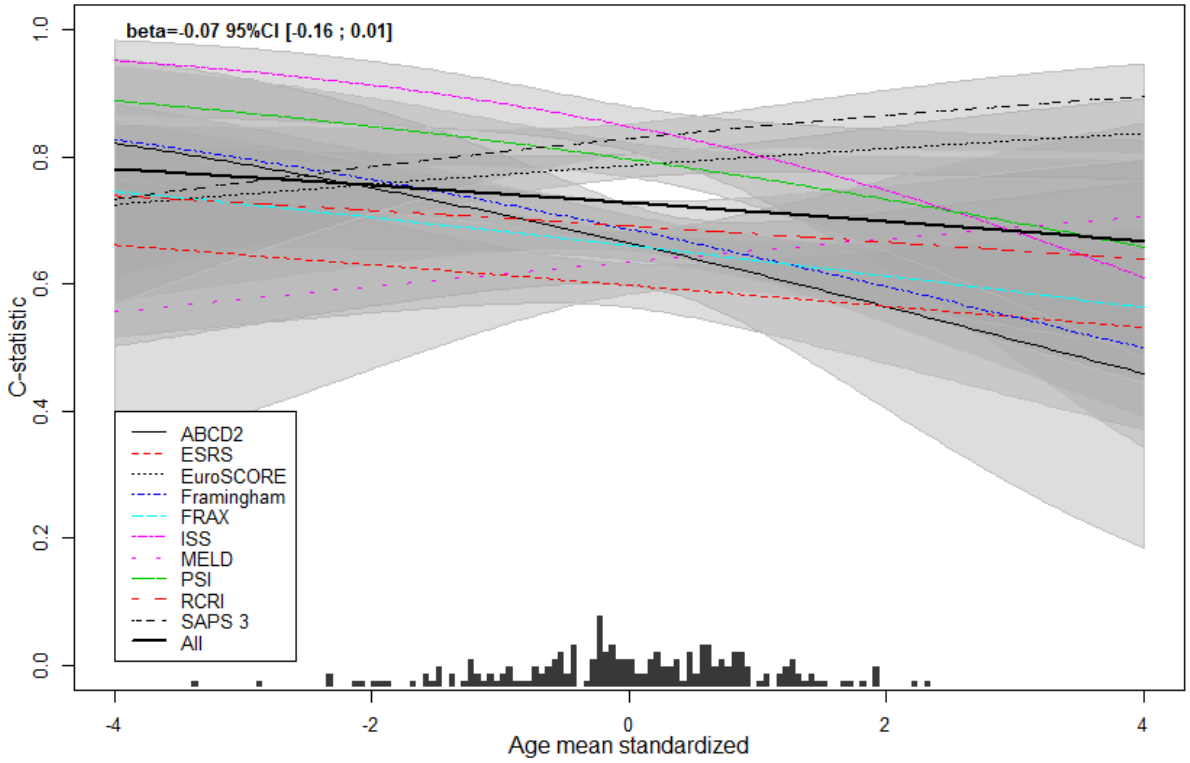



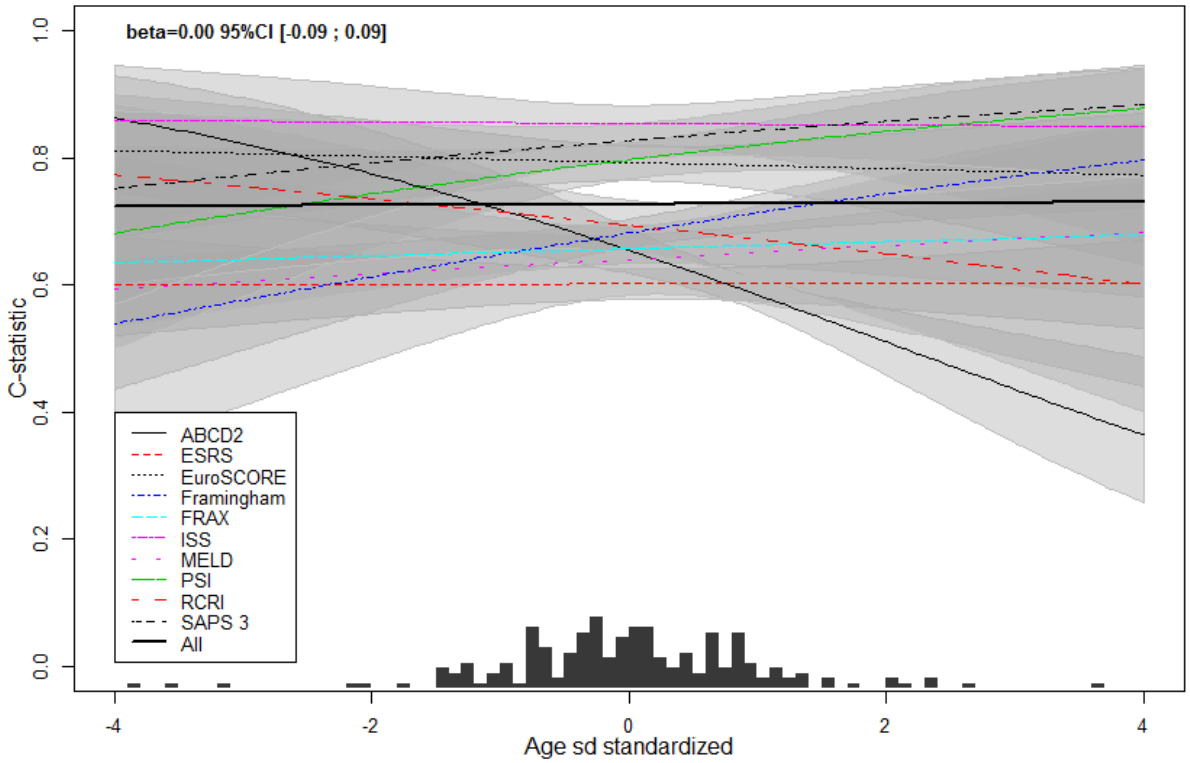
Damen et al. Empirical evidence on the impact of study characteristics and the performance of prediction models: a meta-epidemiological study

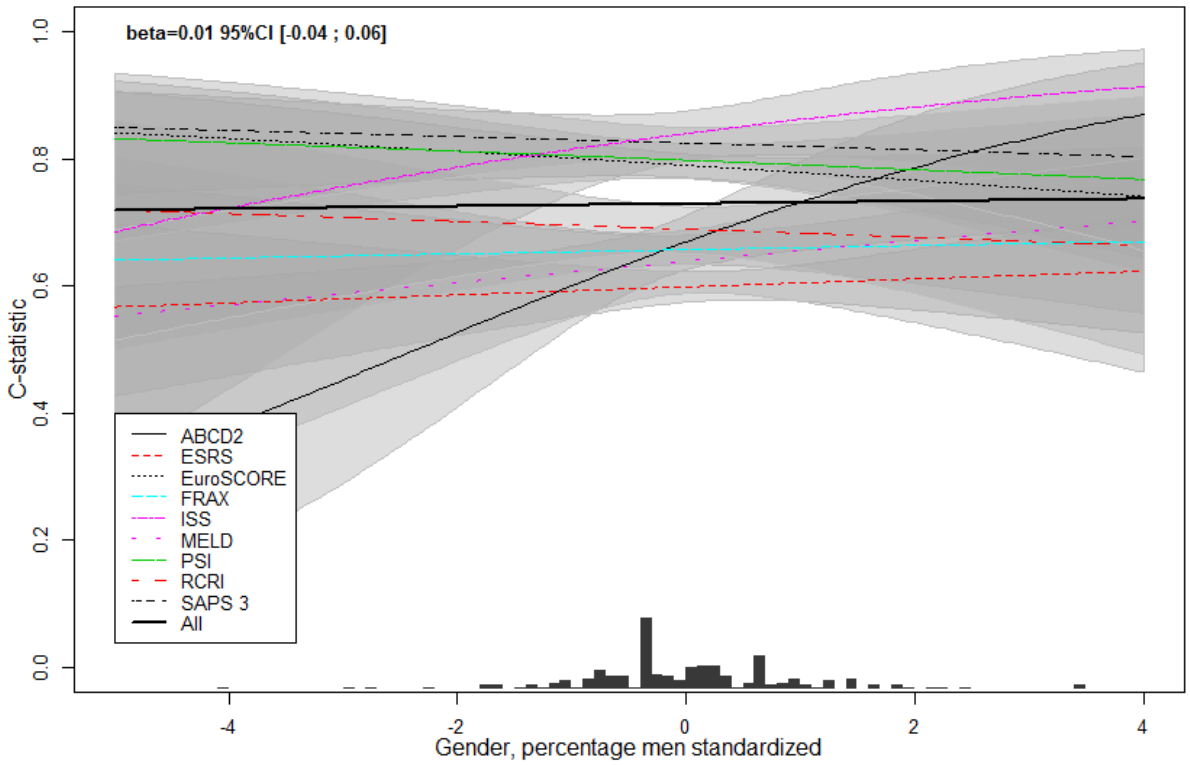


Damen et al. Empirical evidence on the impact of study characteristics and the performance of prediction models: a meta-epidemiological study

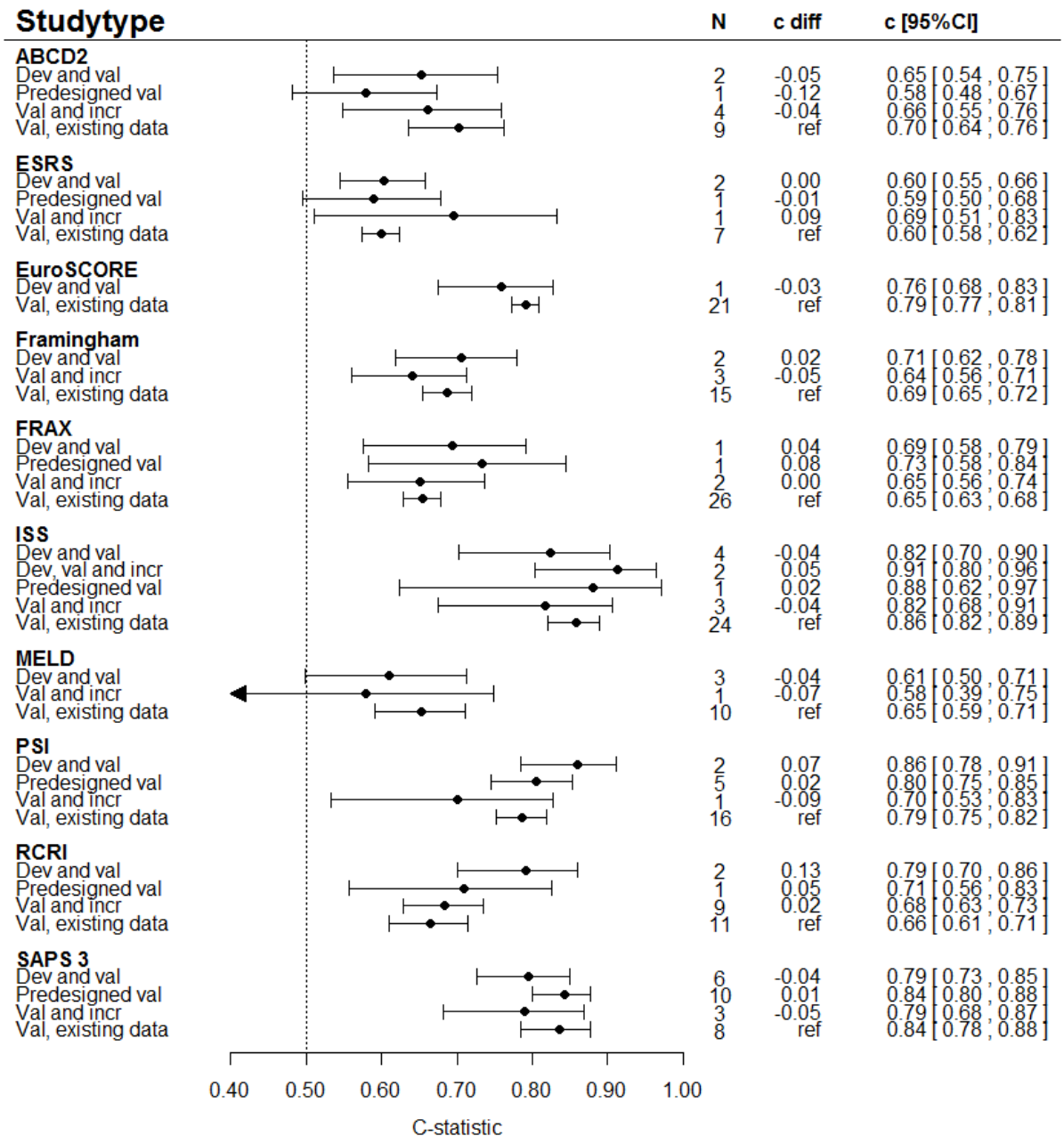


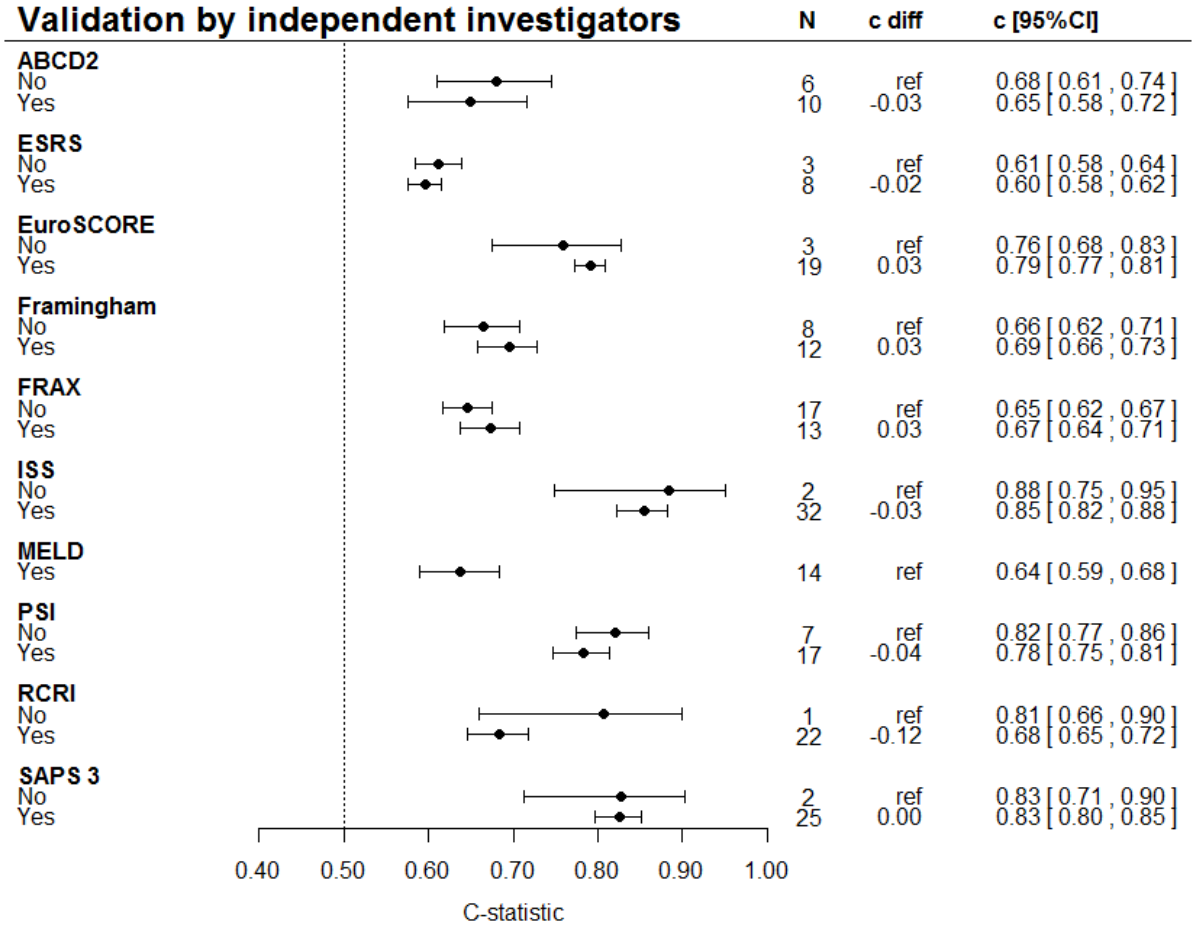



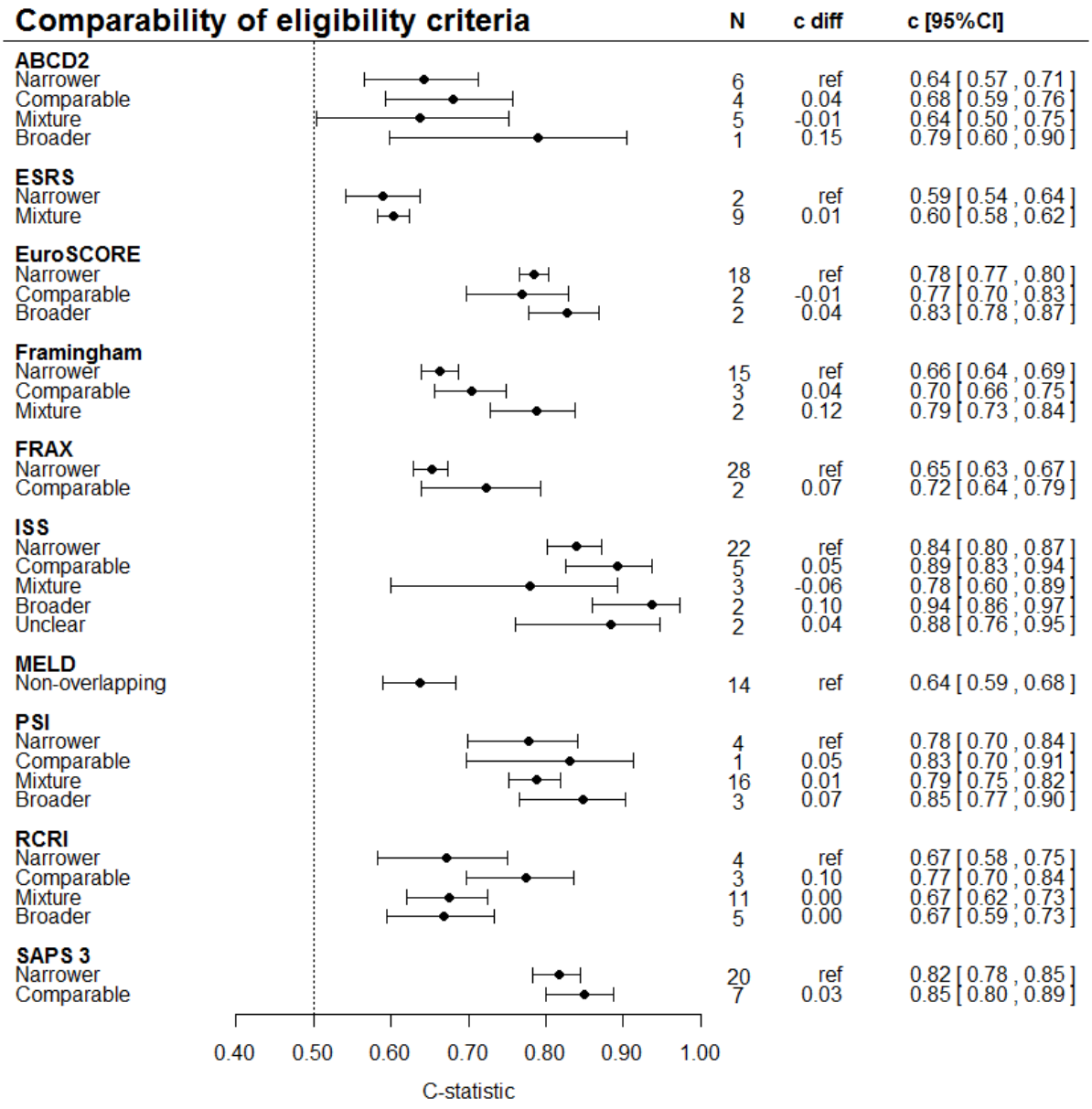


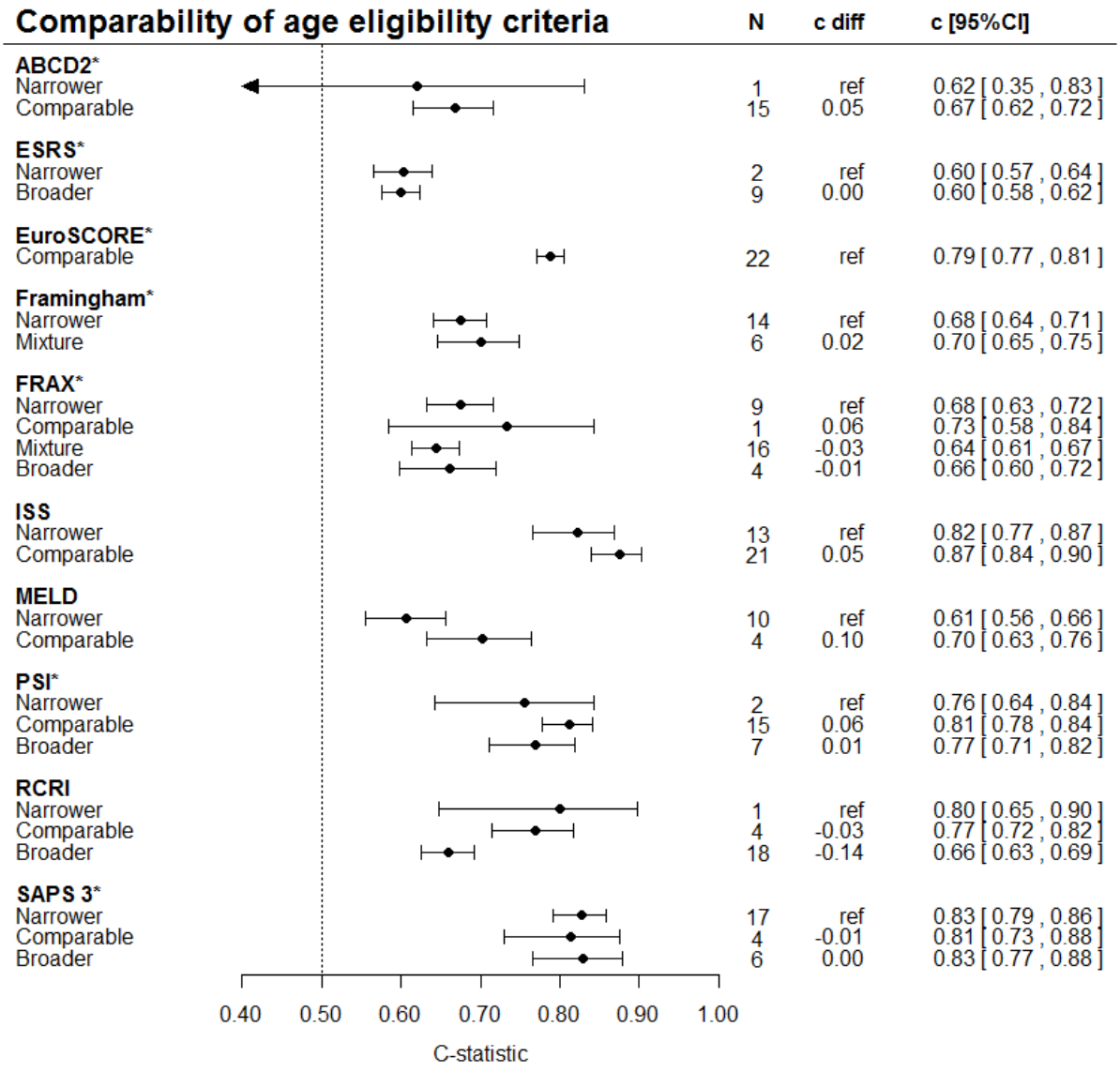

A value of 0 on the x-axis means that the corresponding validation study adopted a common value for that characteristic (i.e. the average value for all validation studies of that specific prediction model), values above 0 mean that values for that characteristic were higher than the average value of all validation studies of that model, whereas values below 0 mean a lower than average value for that characteristic. For example, for the prediction horizon, a value of 0 means that the corresponding validation study adopted the average prediction horizon, values above 0 mean that prediction model performance was assessed for long-term endpoints, whereas values below 0 mean that prediction model performance was assessed for short-term endpoints.

Figure S3: C-statistic in categories of study characteristics within each systematic review

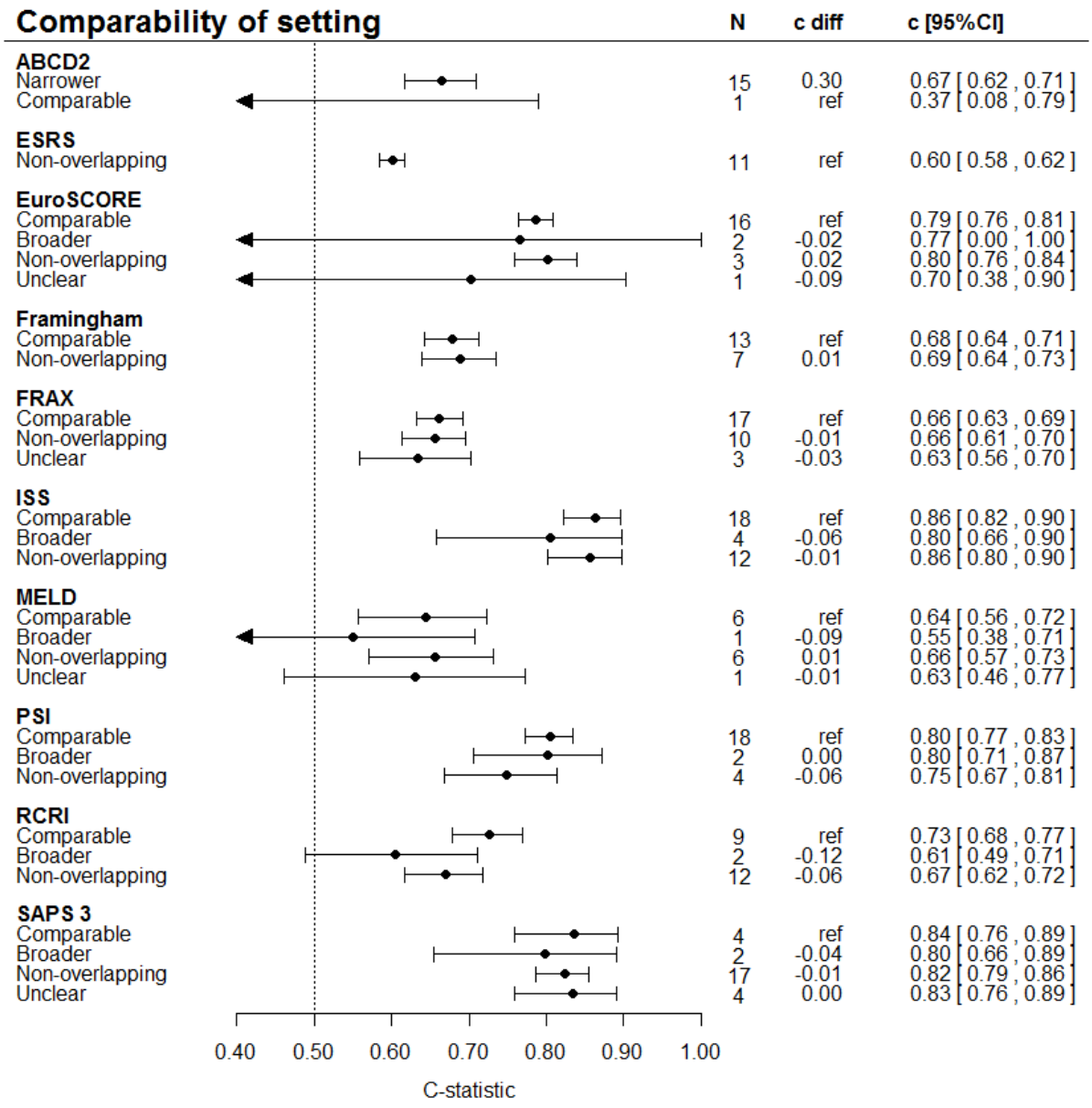


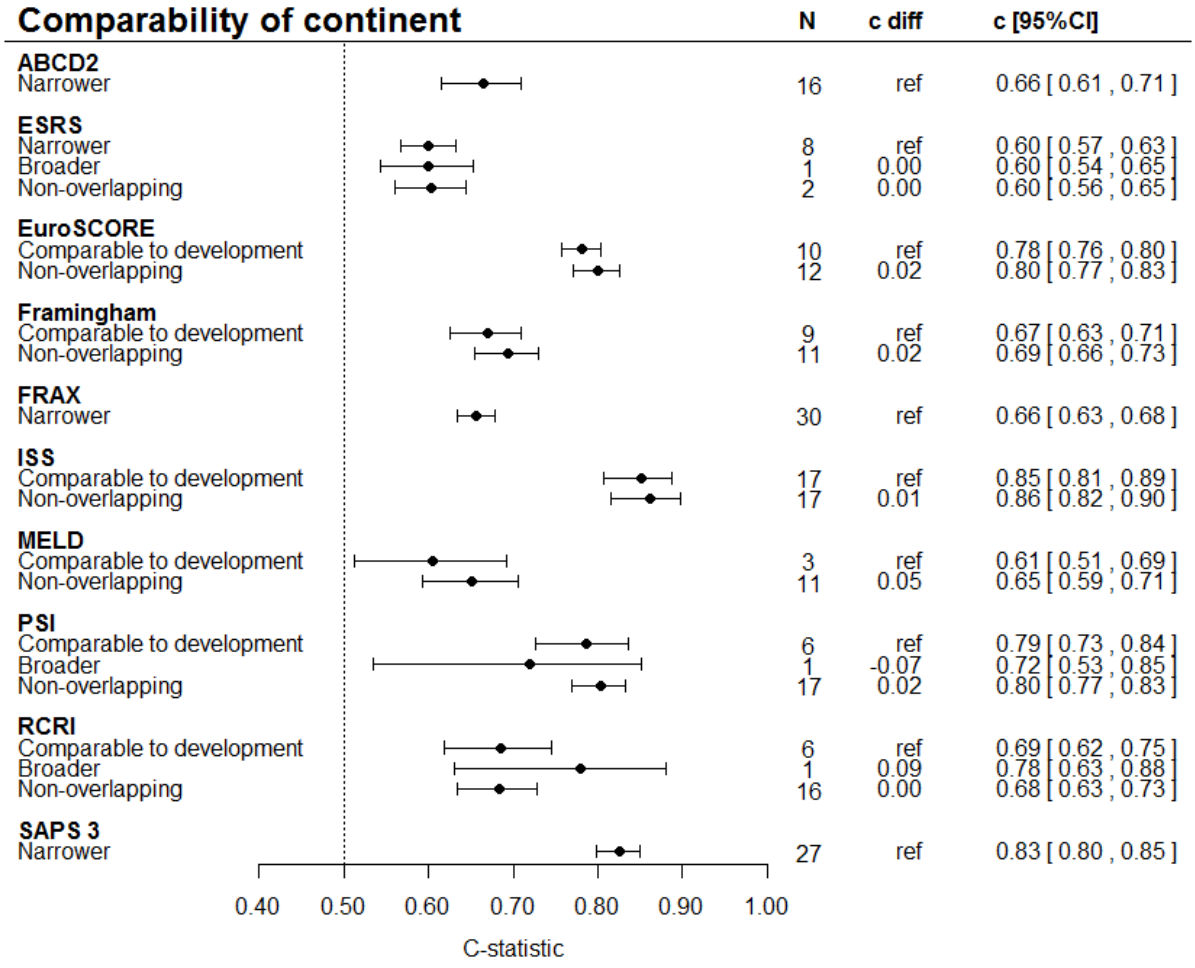

Validation by independent investigators
N
c diff
c [95%CI]
ABCD2
No 6 ref 0.68 [ 0.61 , 0.74 ]
Yes 10 -0.03 0.65 [ 0.58 , 0.72 ]
ESRS
No 3 ref 0.61 [ 0.58 , 0.64 ]
Yes 8 -0.02 0.60 [ 0.58 , 0.62 ]
EuroSCORE
No 3 ref 0.76 [ 0.68 , 0.83 ]
Yes 19 0.03 0.79 [ 0.77 , 0.81 ]
Framingham
No 8 ref 0.66 [ 0.62 , 0.71 ]
Yes 12 0.03 0.69 [ 0.66 , 0.73 ]
FRAX
No 17 ref 0.65 [ 0.62 , 0.67 ]
Yes 13 0.03 0.67 [ 0.64 , 0.71 ]
ISS
No 2 ref 0.88 [ 0.75 , 0.95 ]
Yes 32 -0.03 0.85 [ 0.82 , 0.88 ]
MELD
Yes 14 ref 0.64 [ 0.59 , 0.68 ]
PSI
No 7 ref 0.82 [ 0.77 , 0.86 ]
Yes 17 -0.04 0.78 [ 0.75 , 0.81 ]
RCRI
No 1 ref 0.81 [ 0.66 , 0.90 ]
Yes 22 -0.12 0.68 [ 0.65 , 0.72 ]
SAPS 3
No 2 ref 0.83 [ 0.71 , 0.90 ]
Yes 25 0.00 0.83 [ 0.80 , 0.85 ]
0.40
0.50
0.60
0.70
0.80
0.90
1.00
C-statistic

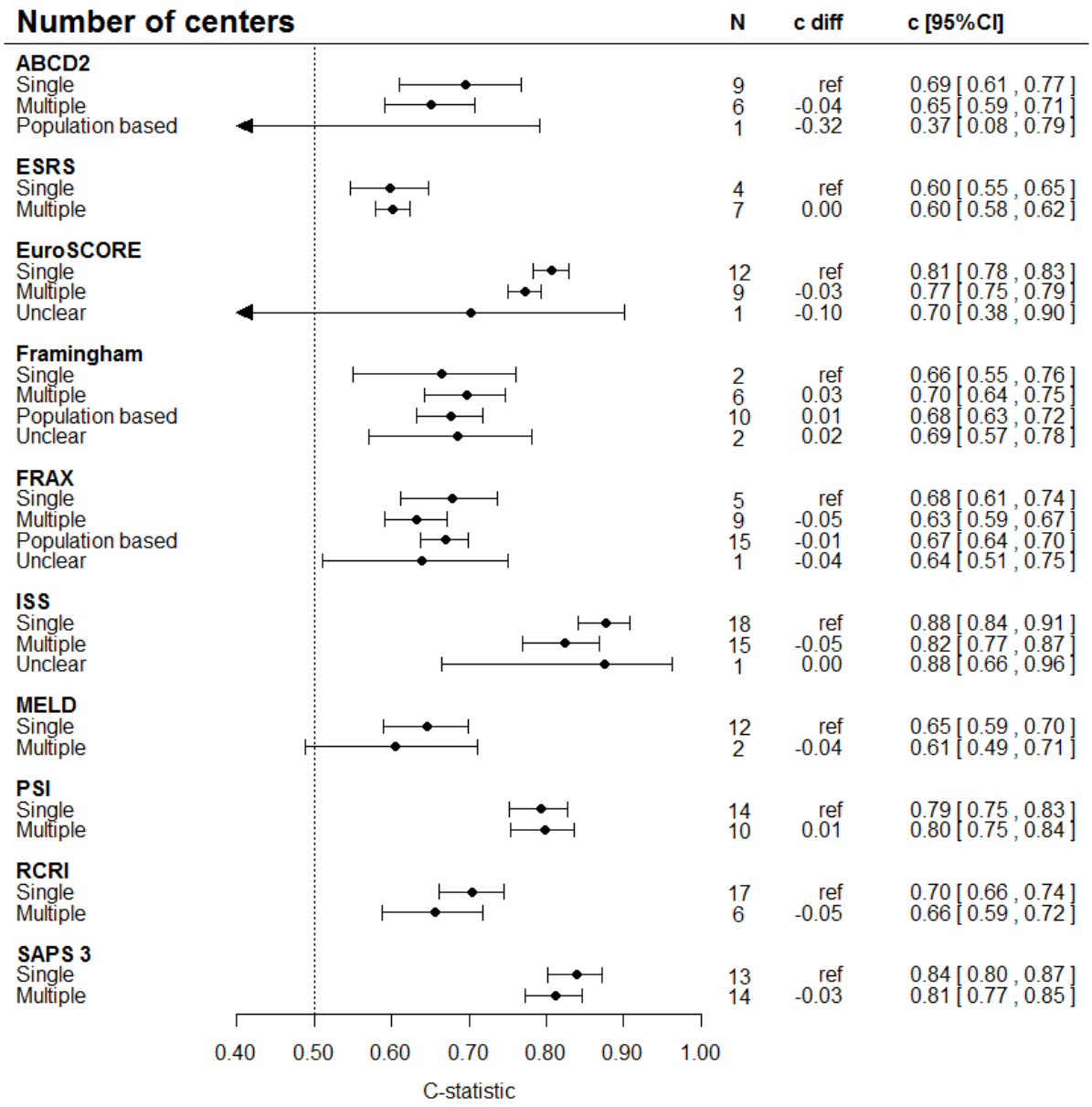


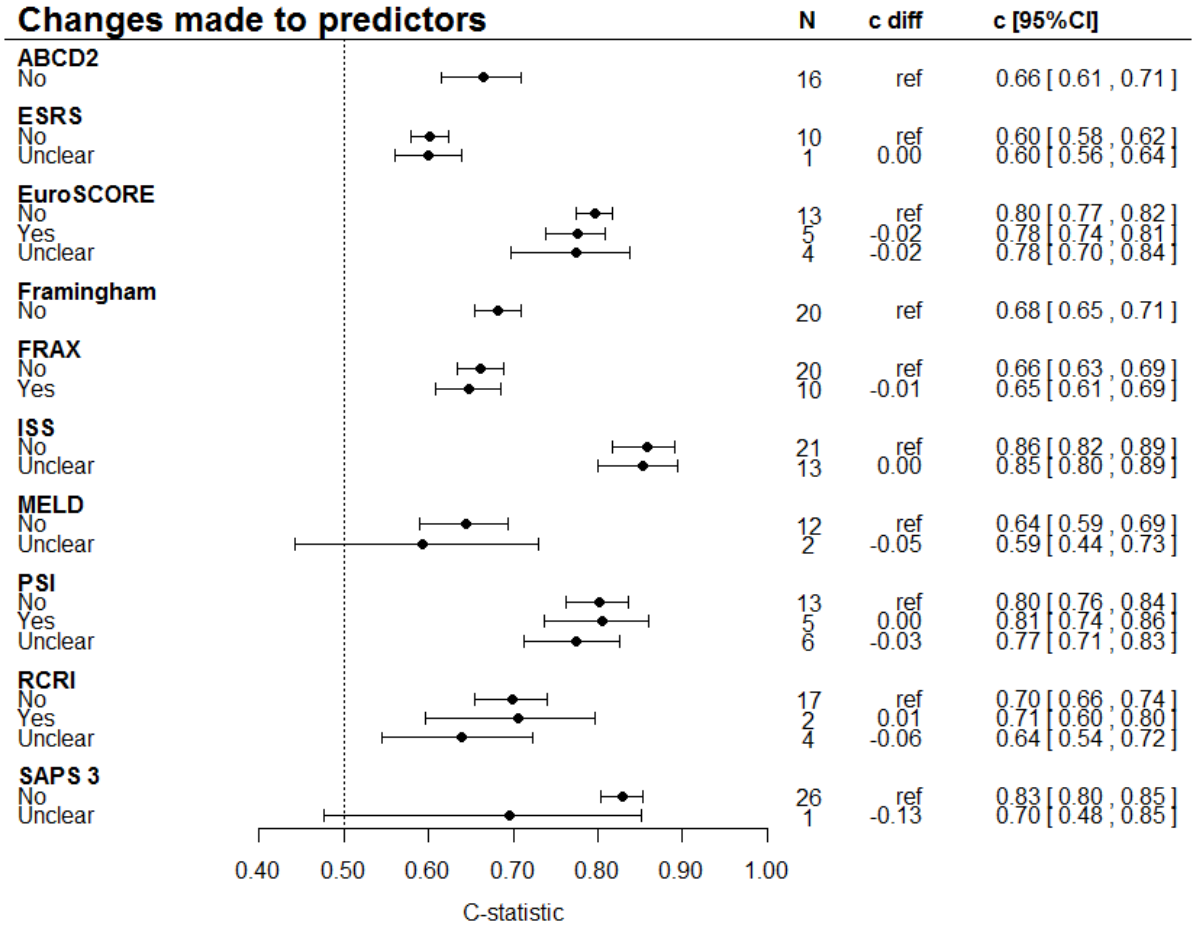


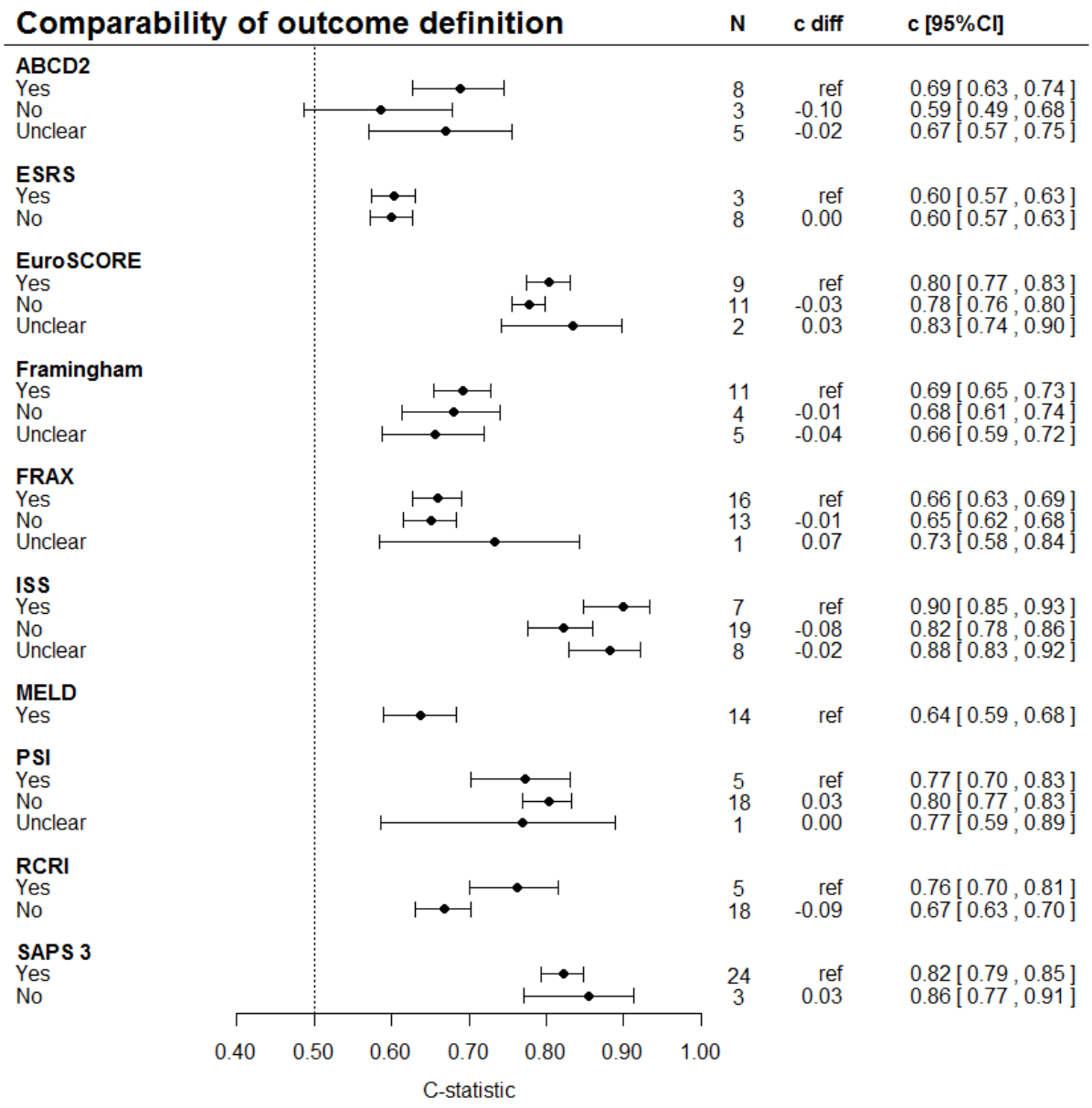

*Models contain age as predictor

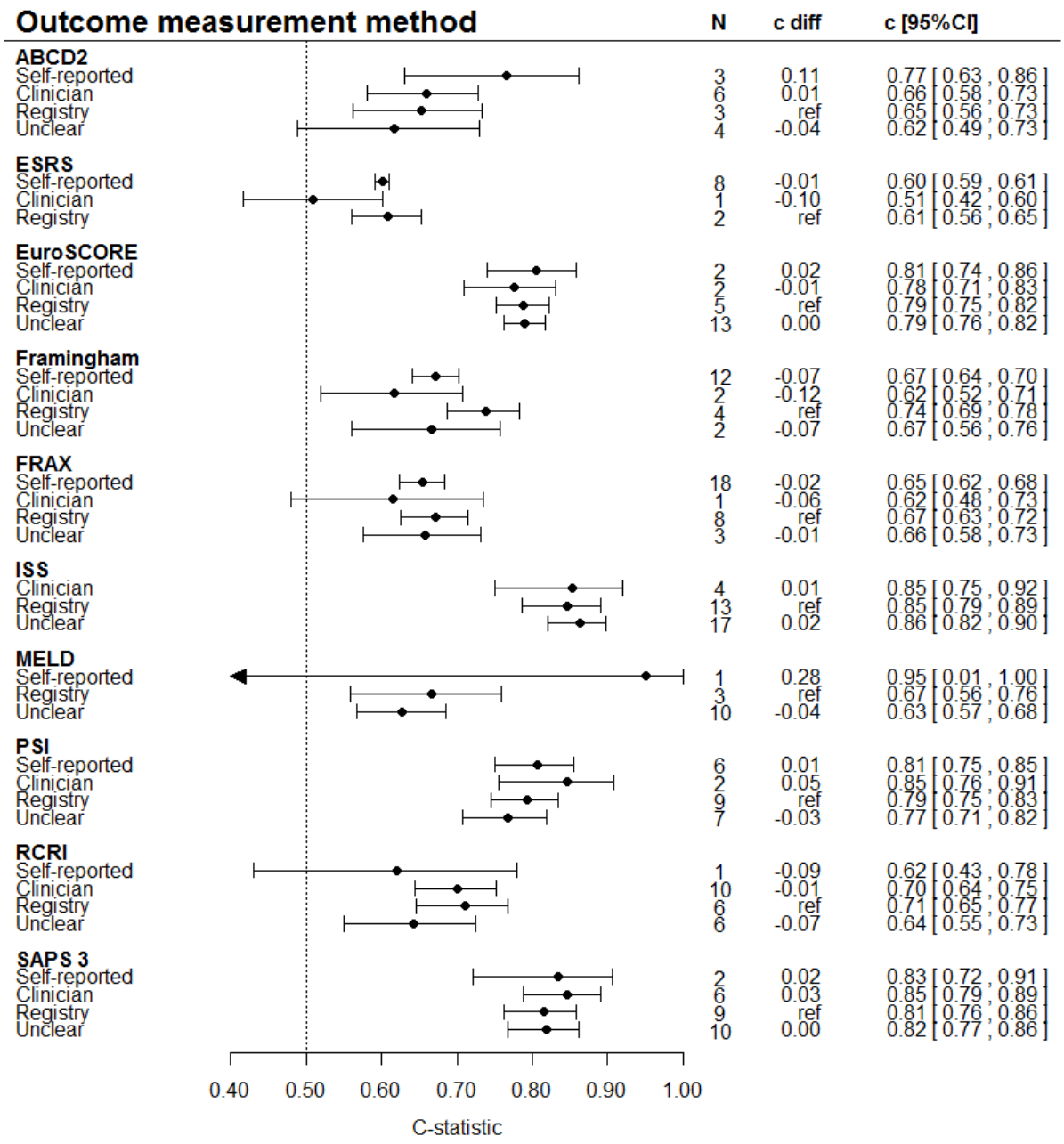


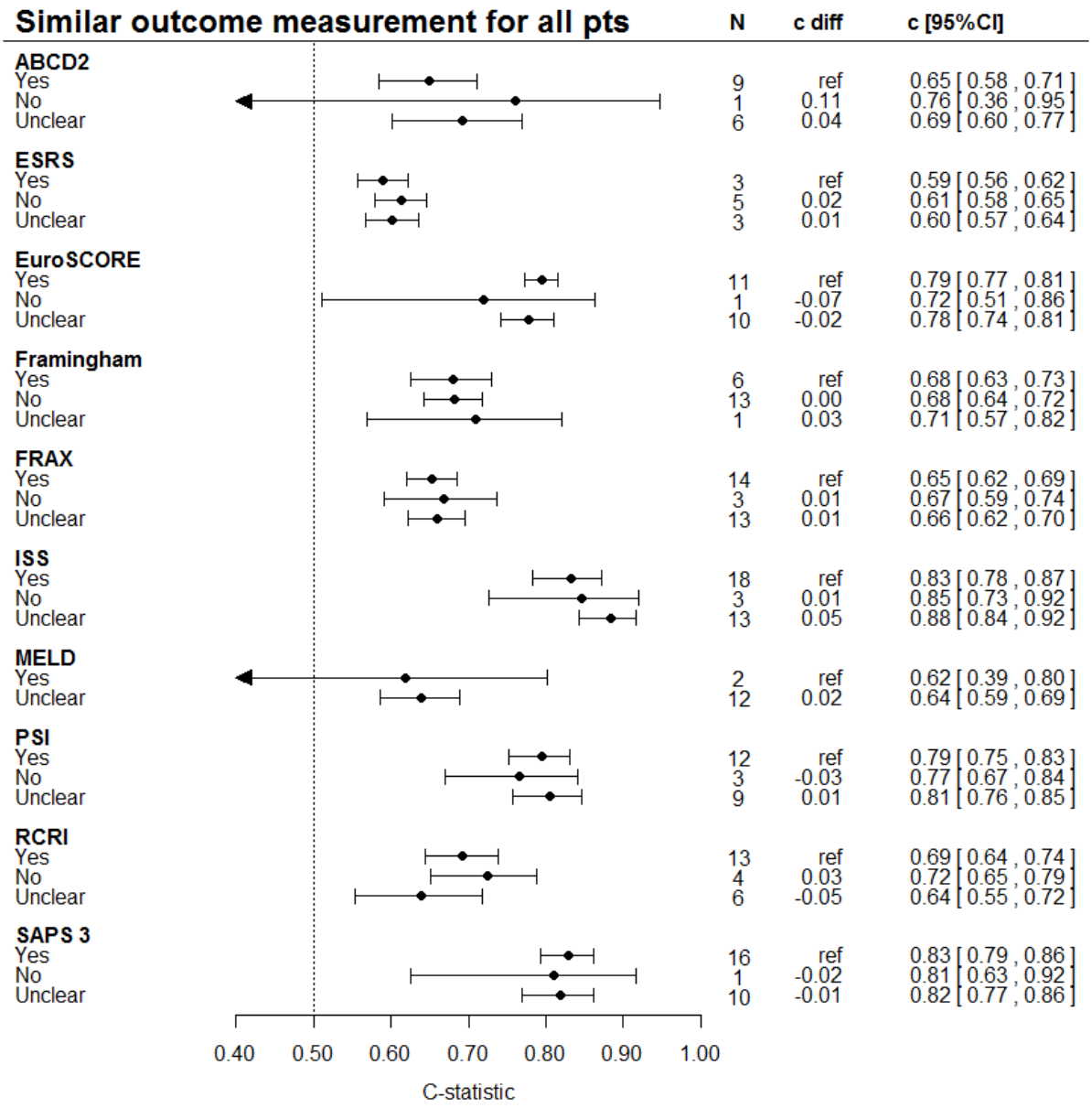


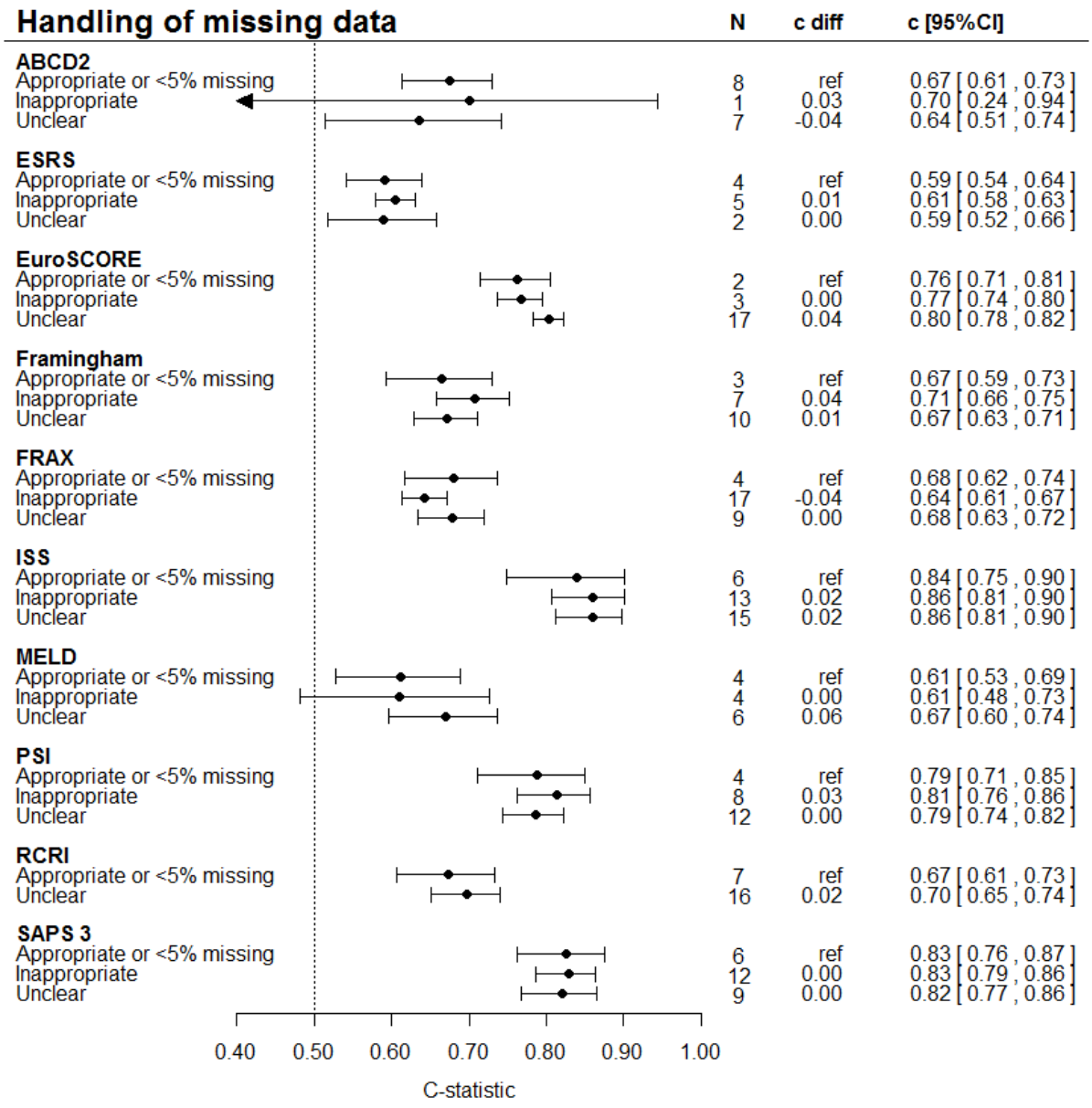



Changes made to predictors
N
c diff
c [95%CI]
ABCD2
No
16
ref
0.66 [ 0.61 , 0.71 ]
ESRS
No
10
ref
0.60 [ 0.58 , 0.62 ]
Unclear
1
0.00
0.60 [ 0.56 , 0.64 ]
EuroSCORE
No
13
ref
0.80 [ 0.77 , 0.82 ]
Yes
5
-0.02
0.78 [ 0.74 , 0.81 ]
Unclear
4
-0.02
0.78 [ 0.70 , 0.84 ]
Framingham
No
20
ref
0.68 [ 0.65 , 0.71 ]
FRAX
No
20
ref
0.66 [ 0.63 , 0.69 ]
Yes
10
-0.01
0.65 [ 0.61 , 0.69 ]
ISS
No
21
ref
0.86 [ 0.82 , 0.89 ]
Unclear
13
0.00
0.85 [ 0.80 , 0.89 ]
MELD
No
12
ref
0.64 [ 0.59 , 0.69 ]
Unclear
2
-0.05
0.59 [ 0.44 , 0.73 ]
PSI
No
13
ref
0.80 [ 0.76 , 0.84 ]
Yes
5
0.00
0.81 [ 0.74 , 0.86 ]
Unclear
6
-0.03
0.77 [ 0.71 , 0.83 ]
RCRI
No
17
ref
0.70 [ 0.66 , 0.74 ]
Yes
2
0.01
0.71 [ 0.60 , 0.80 ]
Unclear
4
-0.06
0.64 [ 0.54 , 0.72 ]
SAPS 3
No
26
ref
0.83 [ 0.80 , 0.85 ]
Unclear
1
-0.13
0.70 [ 0.48 , 0.85 ]
0.40
0.50
0.60
0.70
0.80
0.90
1.00
C-statistic










C-statistic for categories of study characteristics, pooled using univariable meta-regression analyses per systematic review. N represents the number of external validation studies in a specific category. C diff represents the difference in c-statistic with regard to a reference category (indicated with 'ref'). Dev: development, val: validation, incr: incremental value, pts: patients.

Figure S4: Associations between categorical variables and total OE ratio

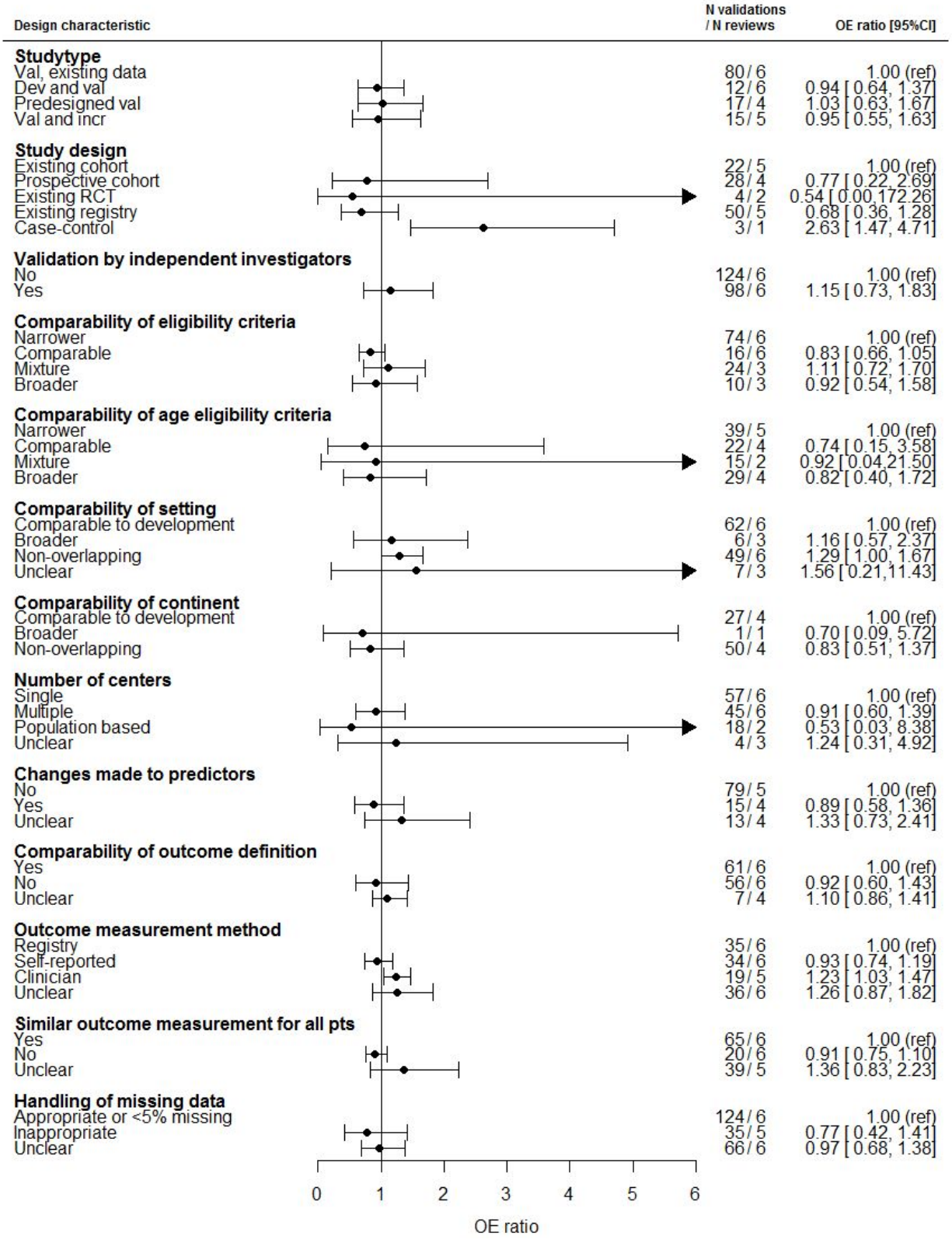
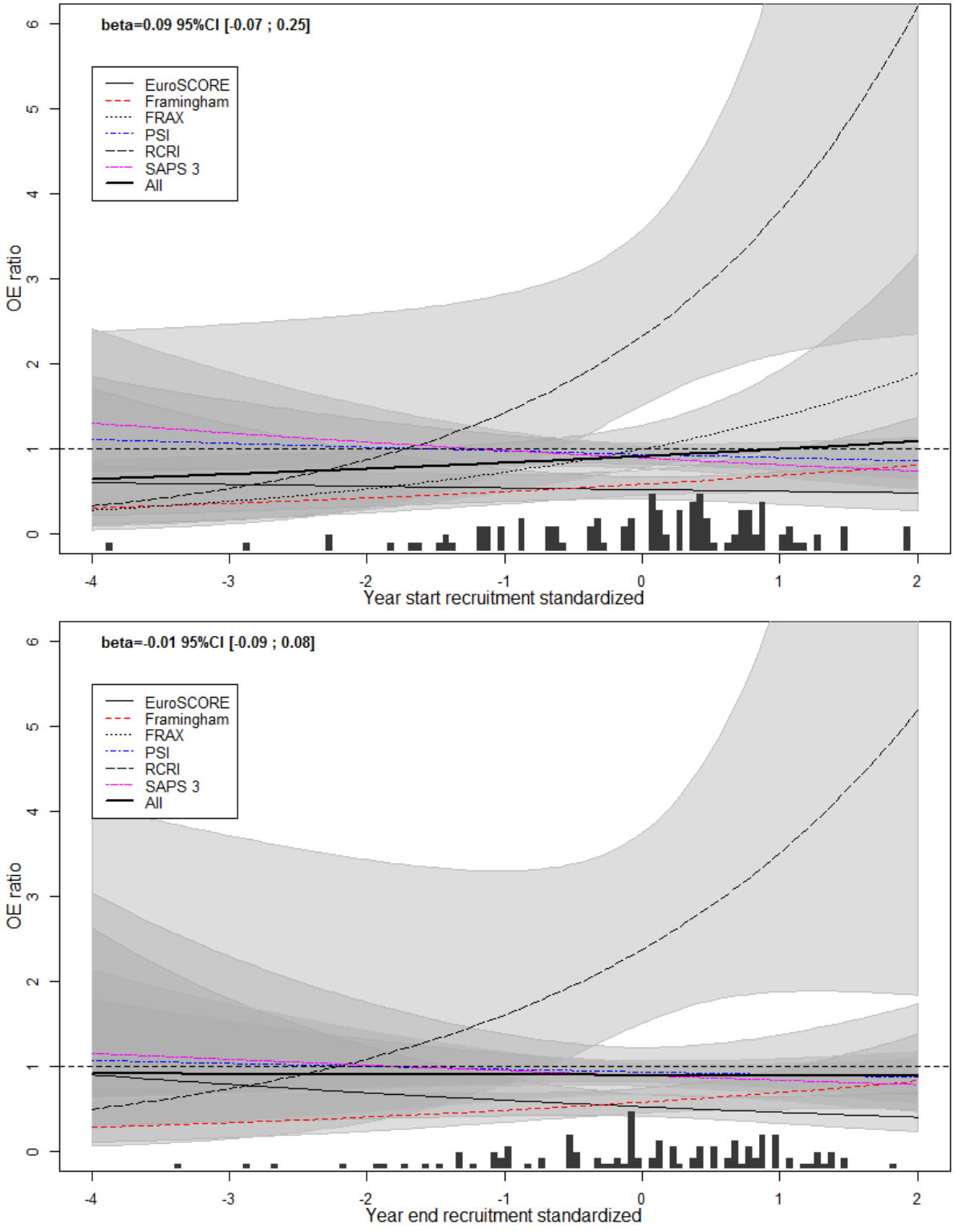
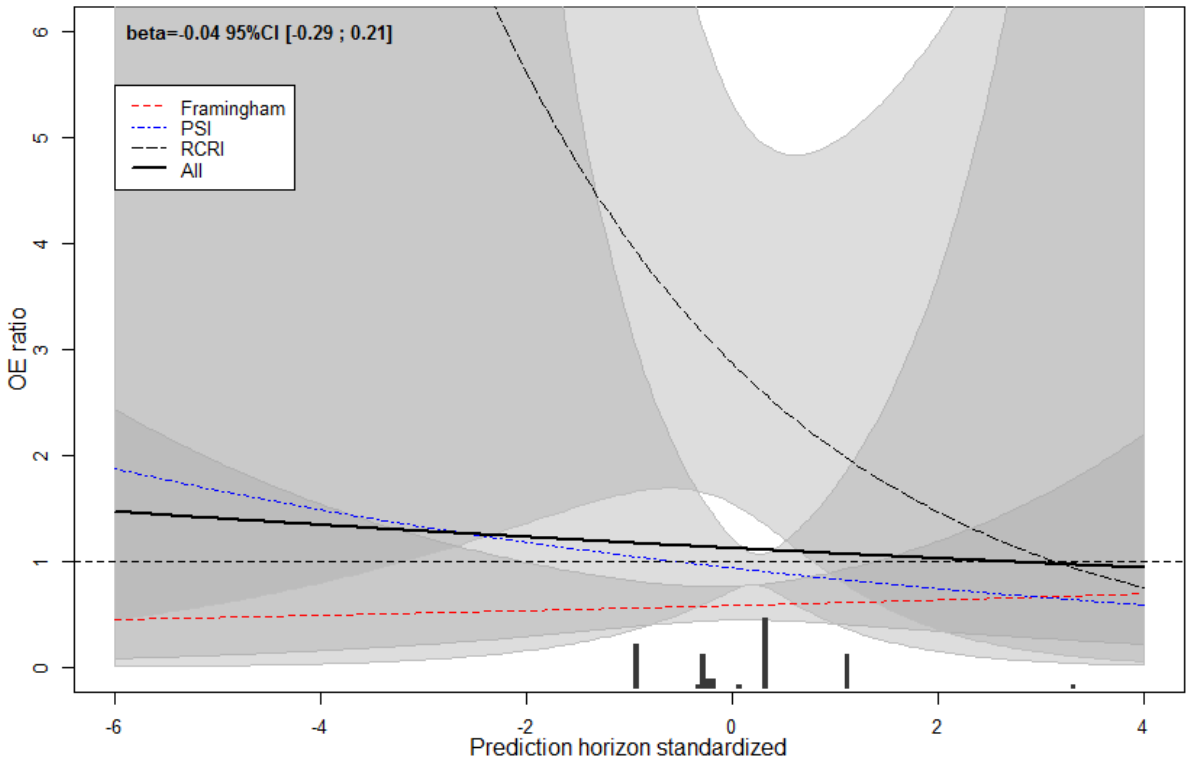
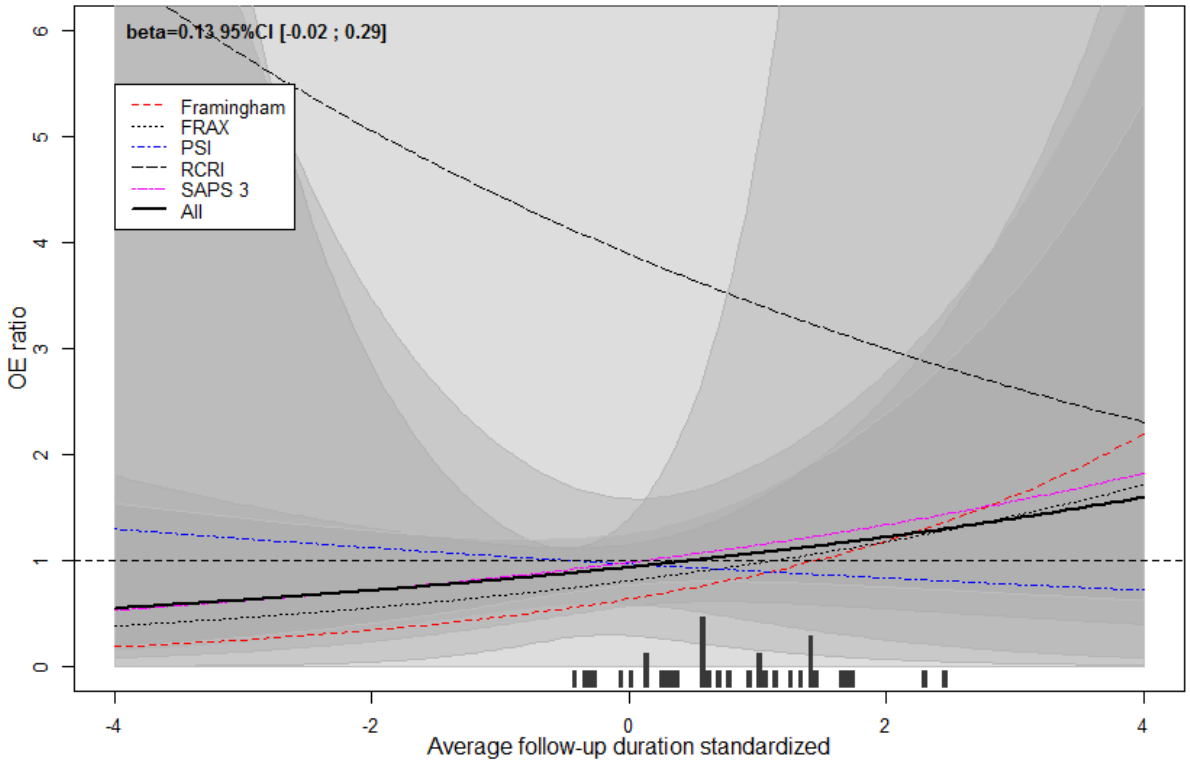


Figure S5: Associations between continuous variables and total OE ratio

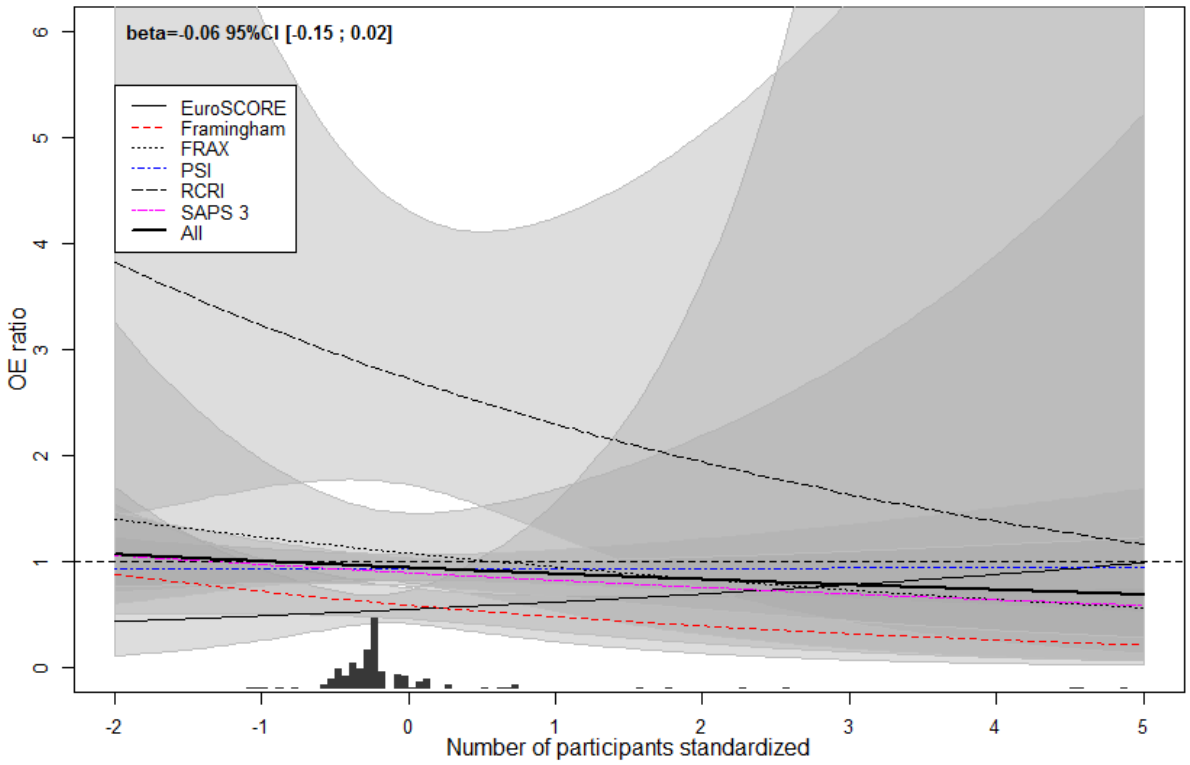


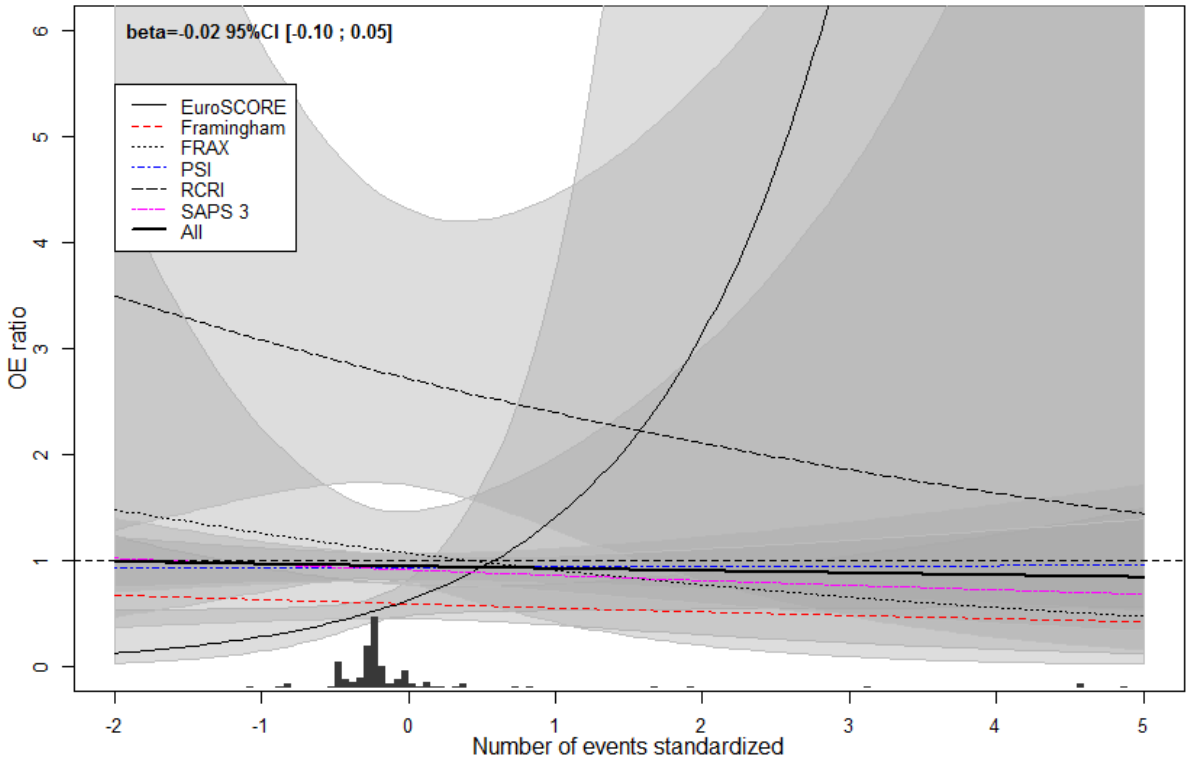
Damen et al. Empirical evidence on the impact of study characteristics and the performance of prediction models: a meta-epidemiological study



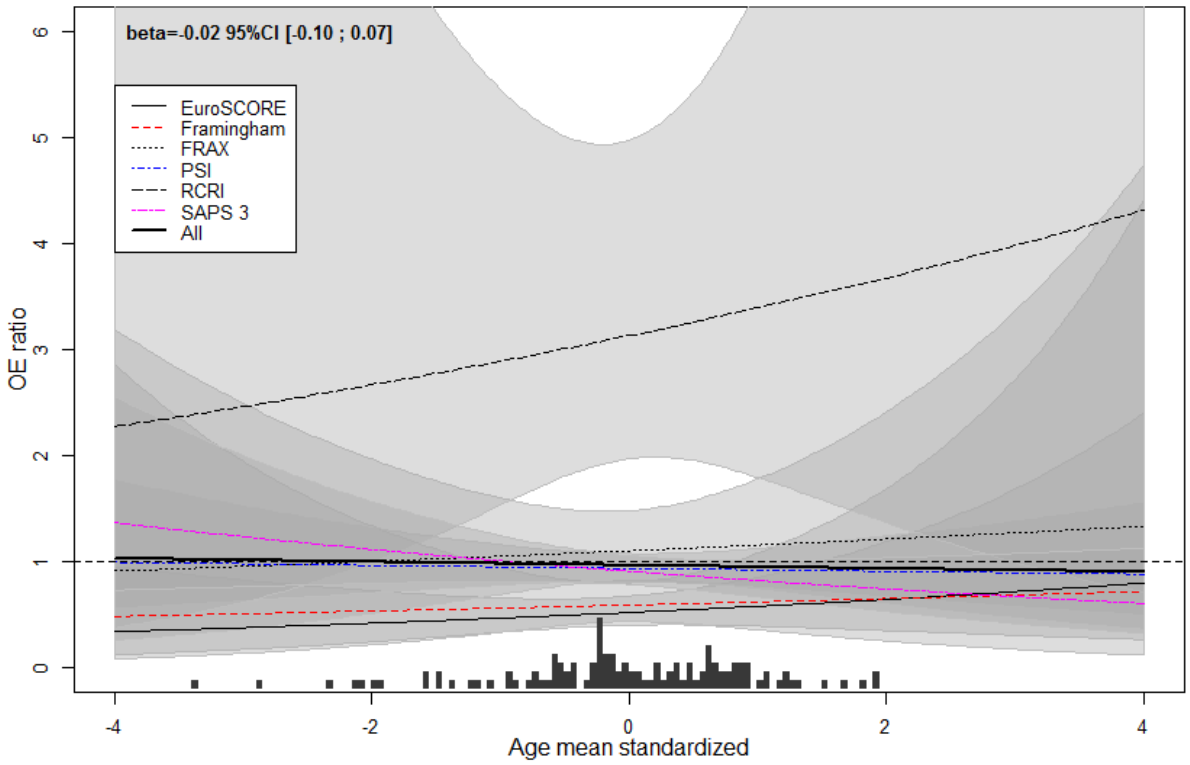



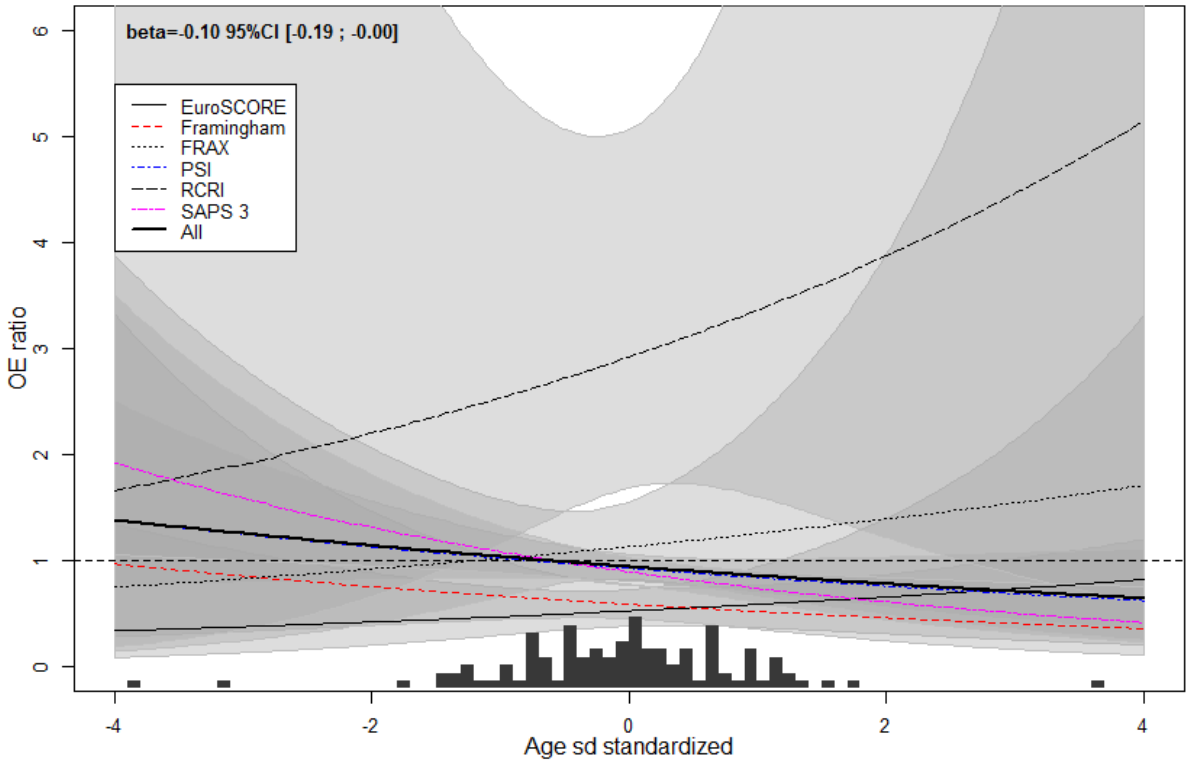
Damen et al. Empirical evidence on the impact of study characteristics and the performance of prediction models: a meta-epidemiological study

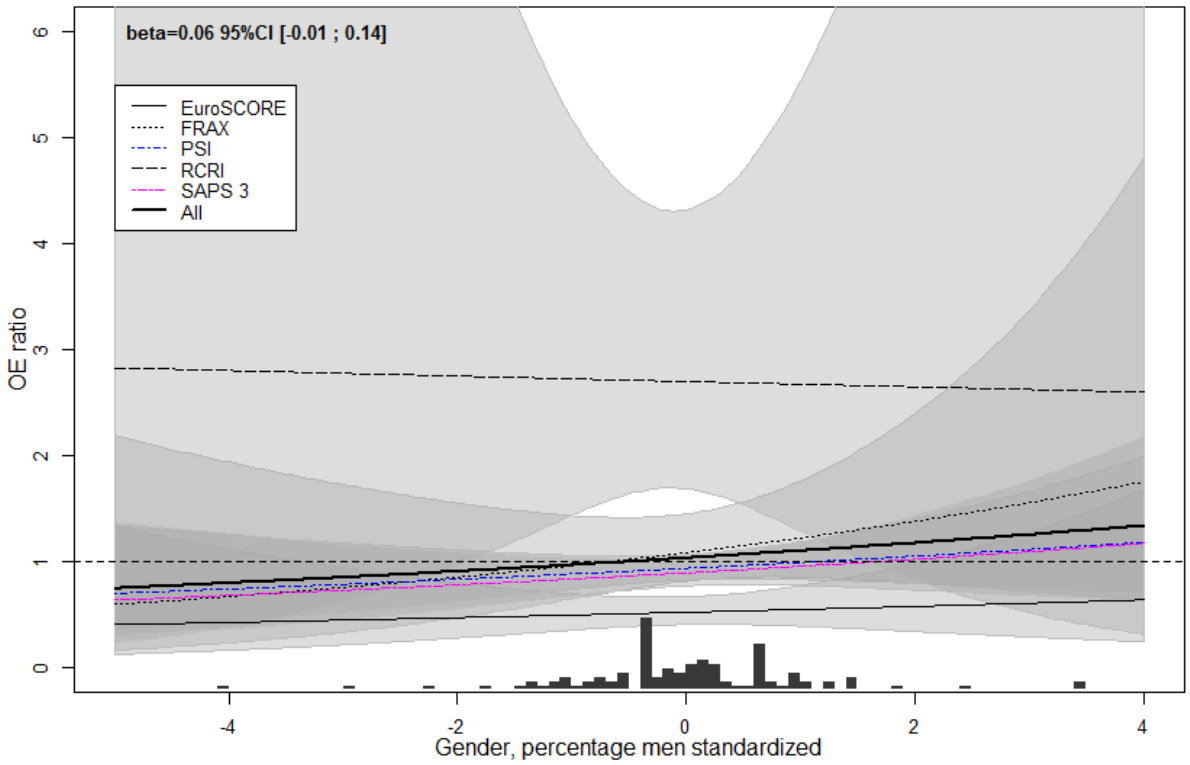


Damen et al. Empirical evidence on the impact of study characteristics and the performance of prediction models: a meta-epidemiological study

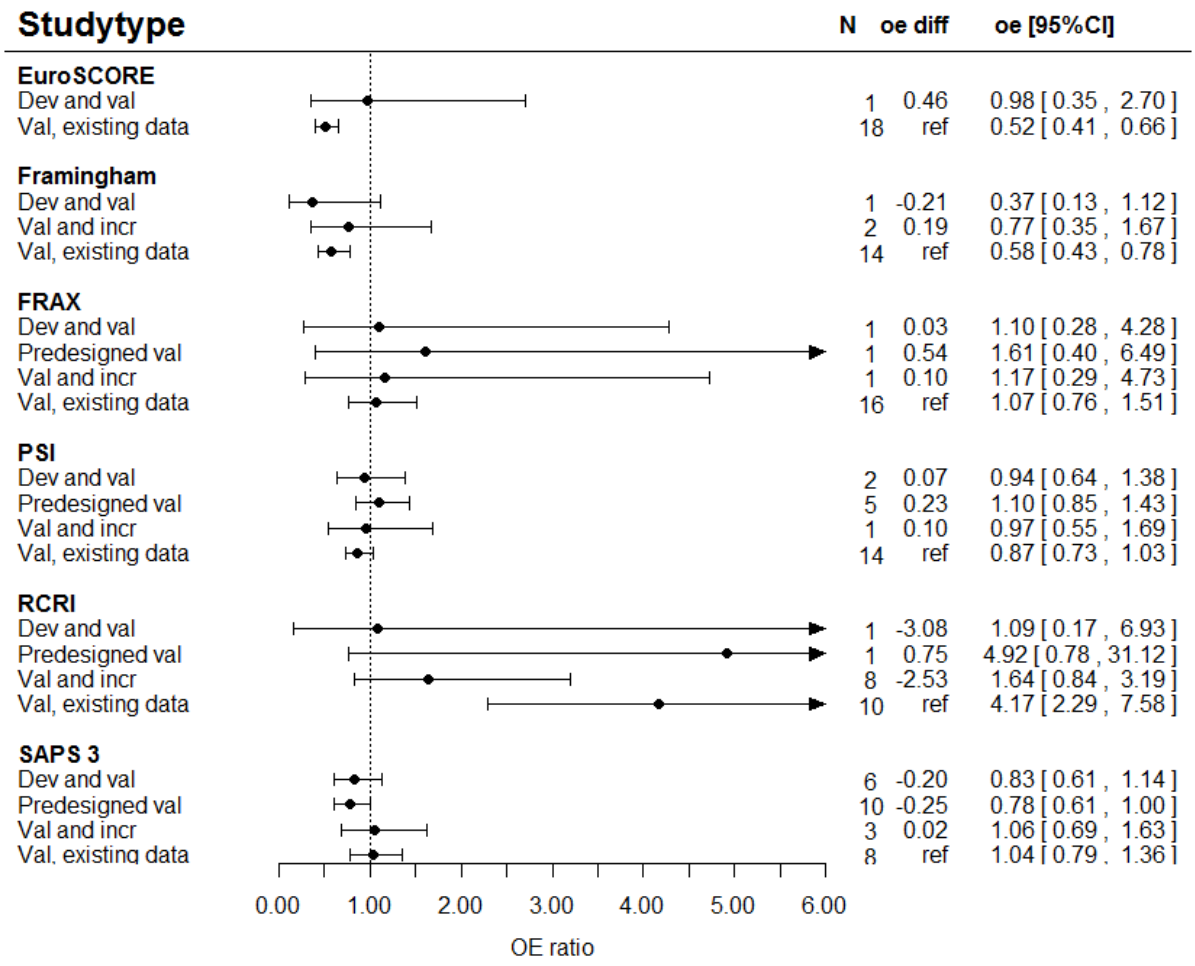


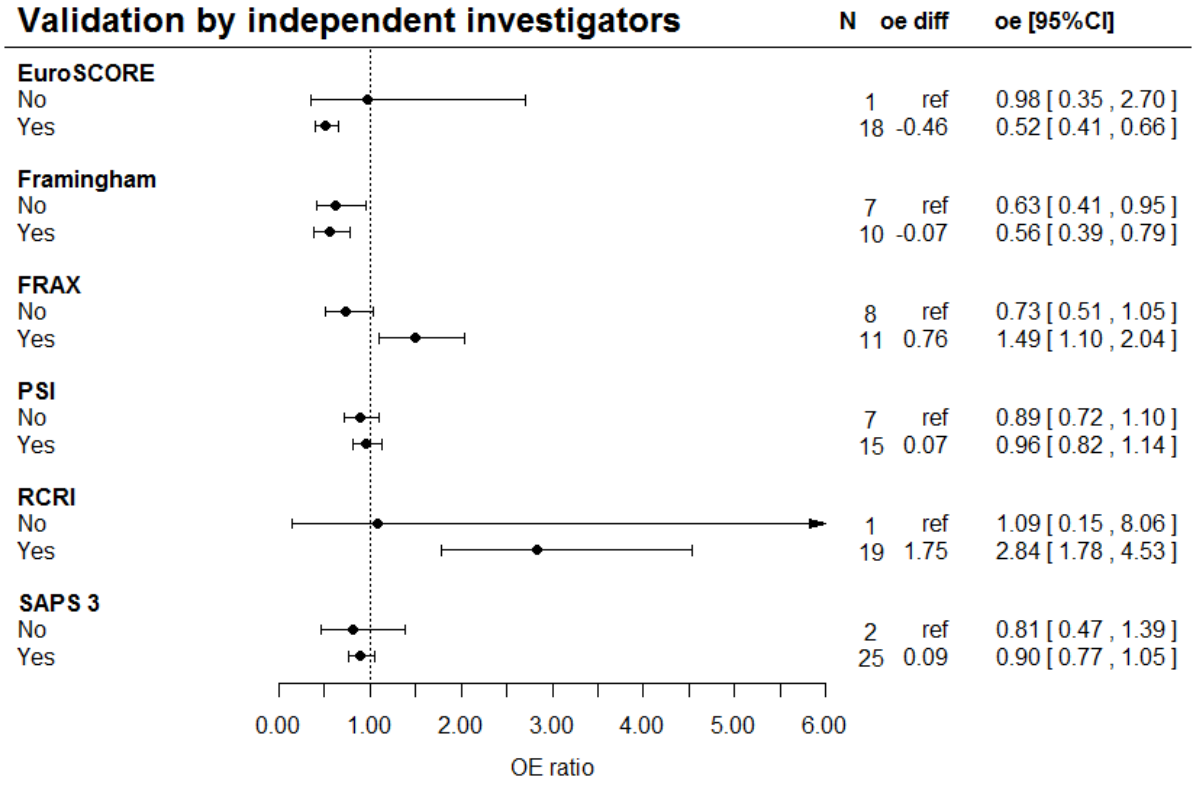



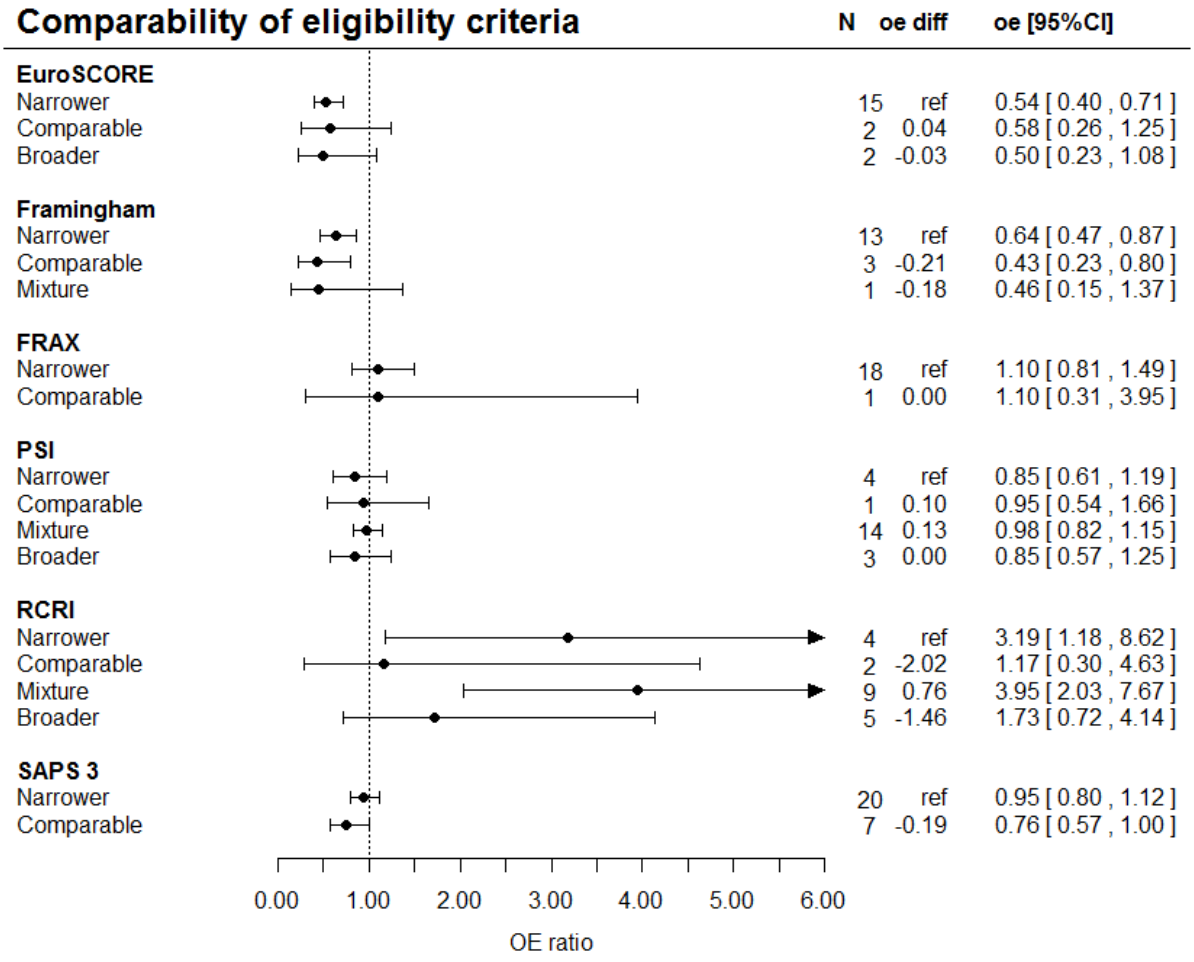


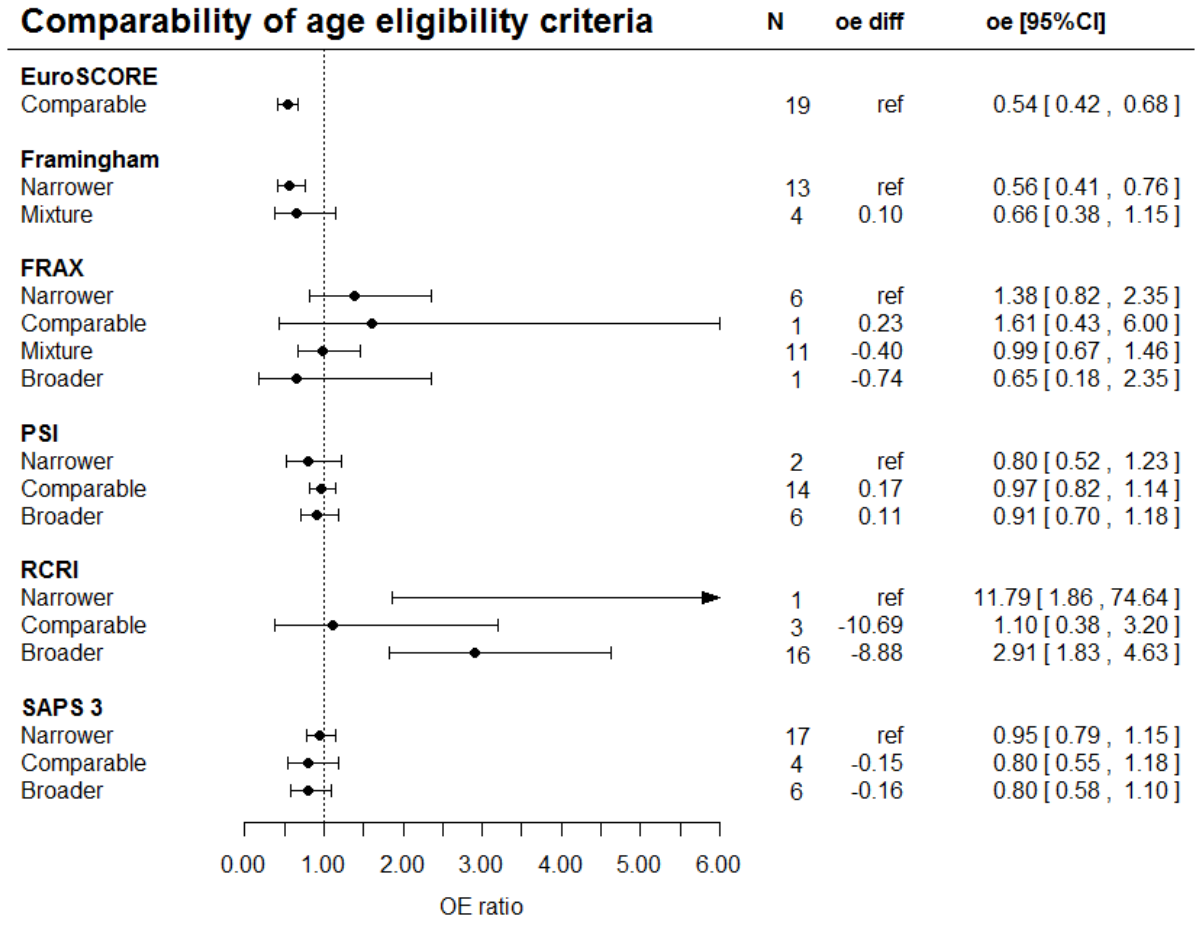

A value of 0 on the x-axis means that the corresponding validation study adopted a common value for that characteristic (i.e. the average value for all validation studies of that specific prediction model), values above 0 mean that values for that characteristic were higher than the average value of all validation studies of that model, whereas values below 0 mean a lower than average value for that characteristic. For example, for the prediction horizon, a value of 0 means that the corresponding validation study adopted the average prediction horizon, values above 0 mean that prediction model performance was assessed for long-term endpoints, whereas values below 0 mean that prediction model performance was assessed for short-term endpoints.

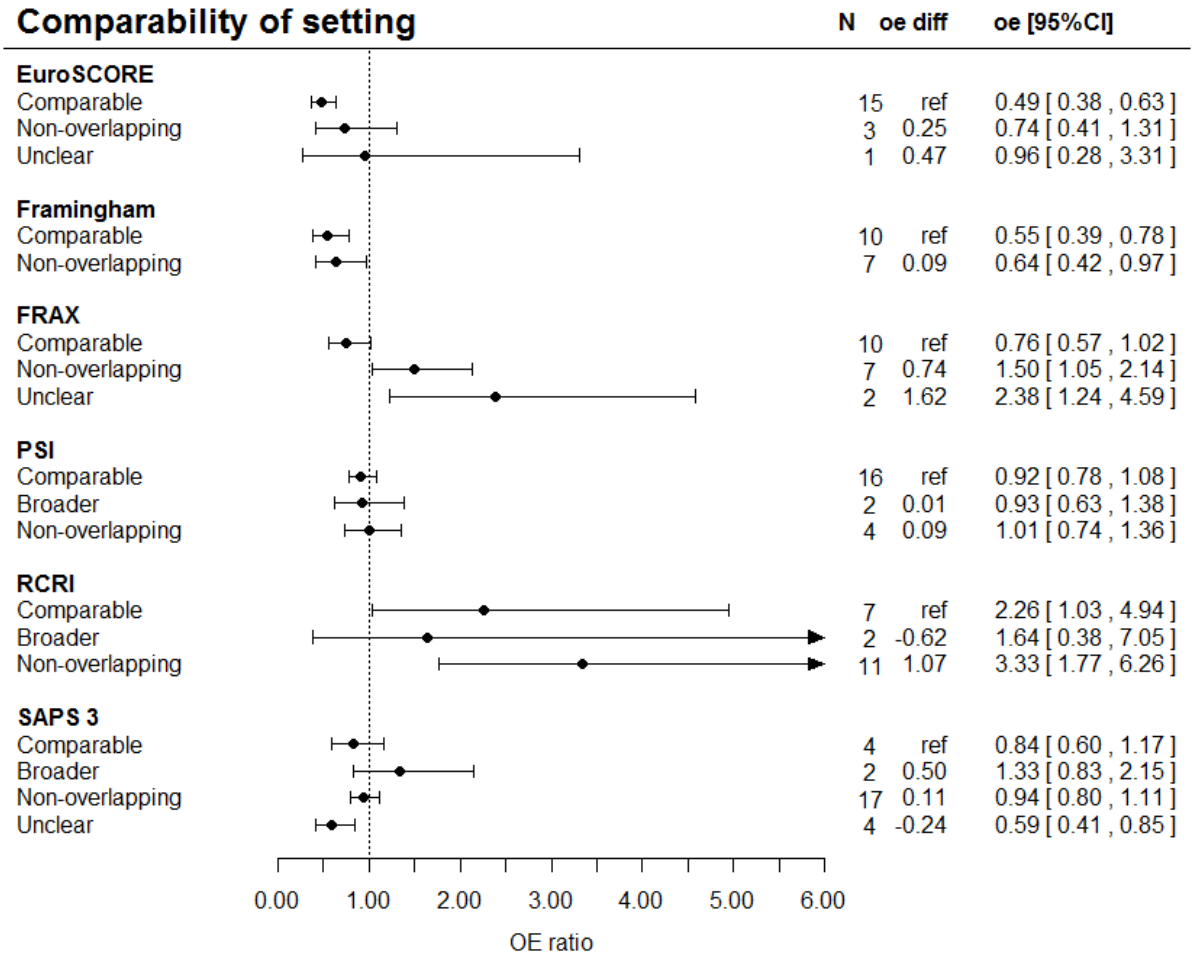
Figure S6: Total OE ratio in categories of study characteristics within each systematic review

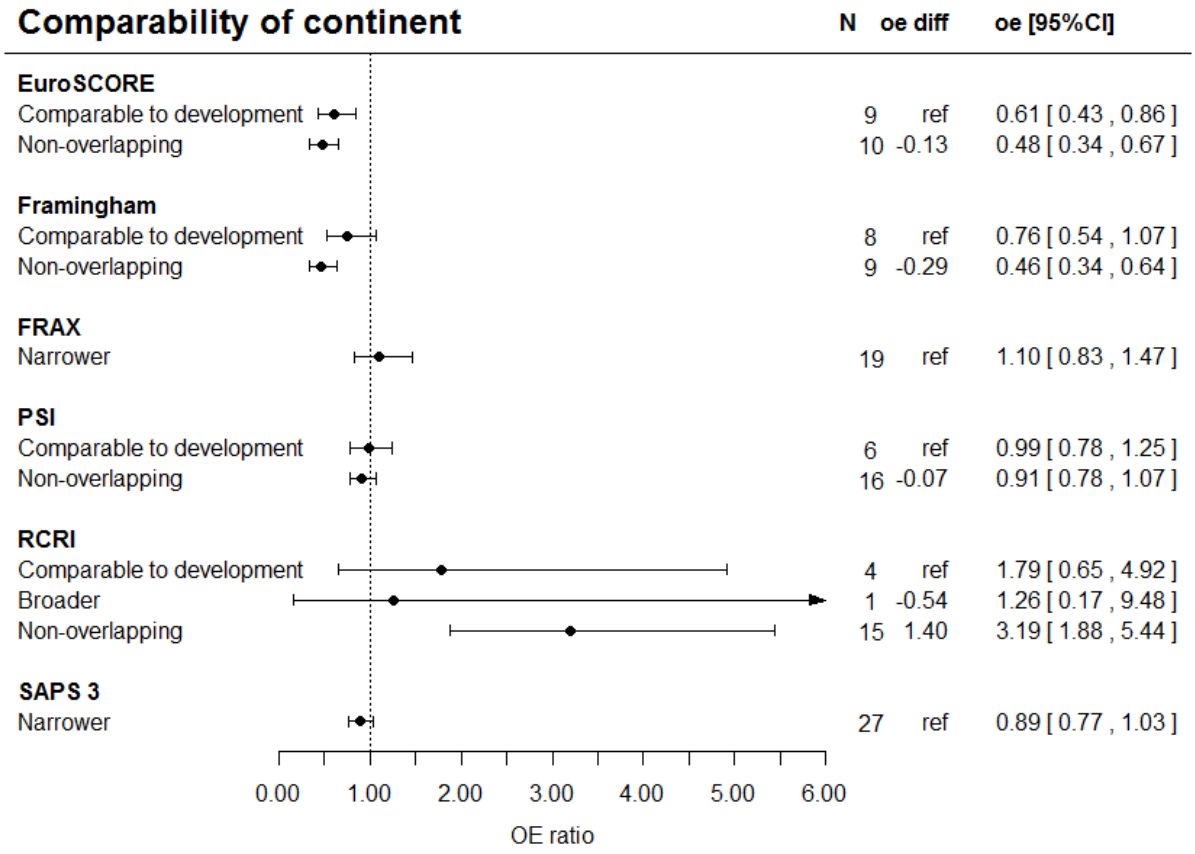


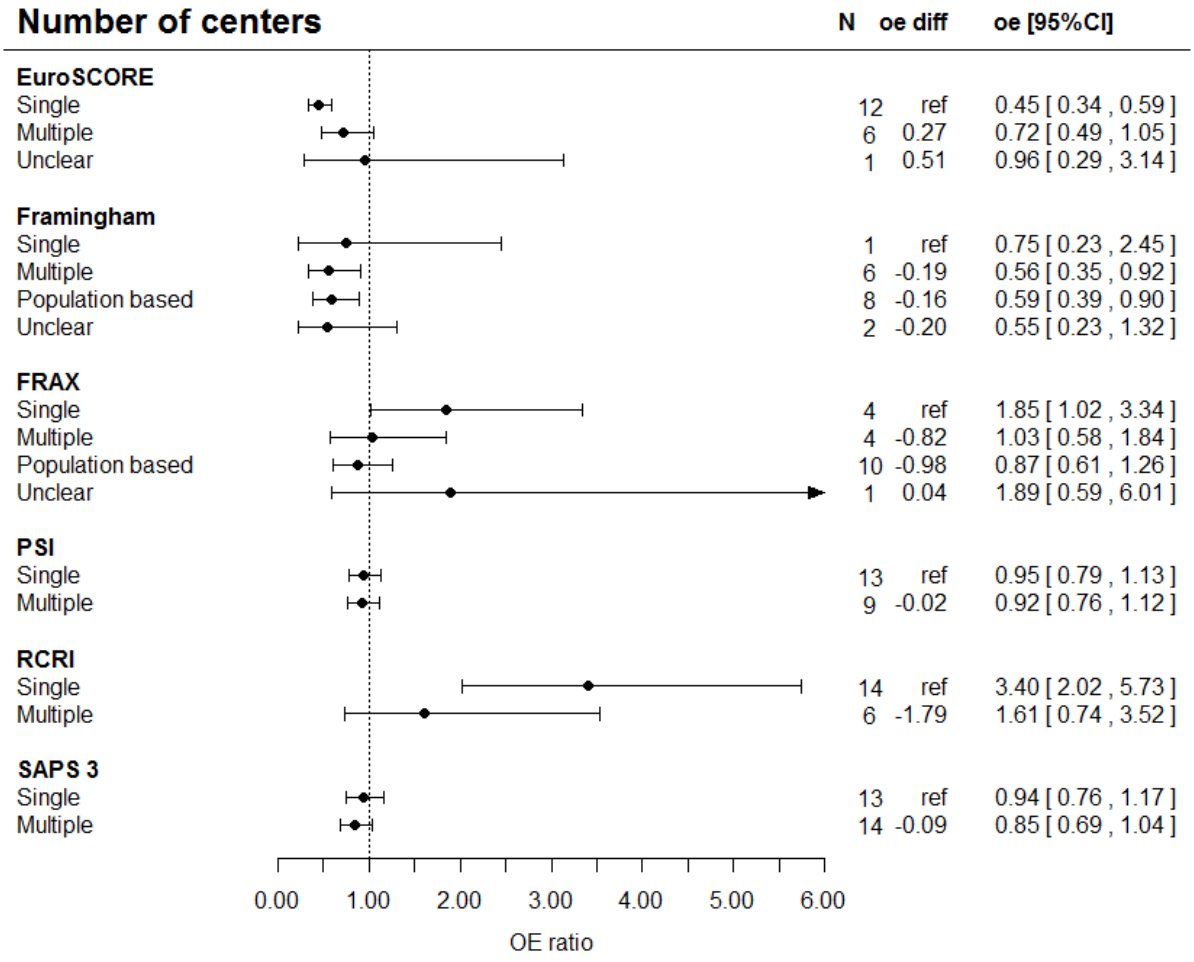


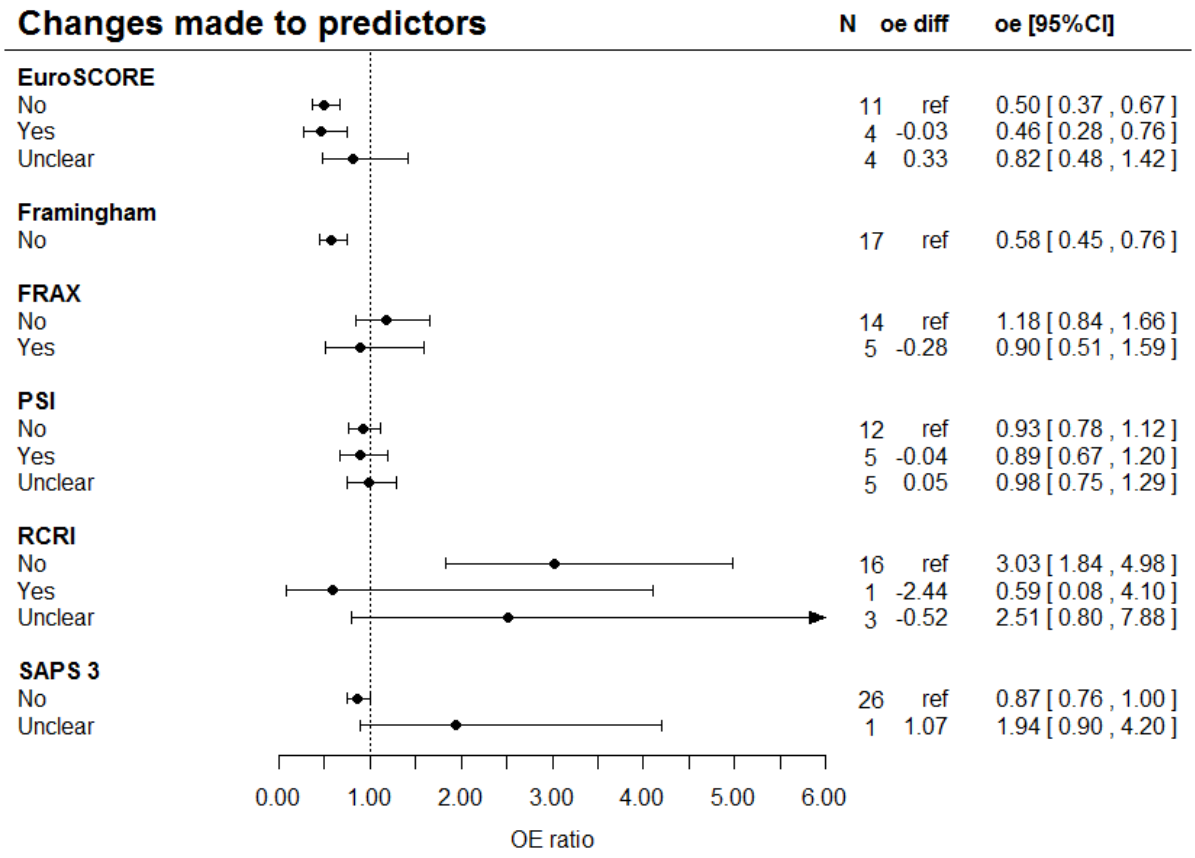


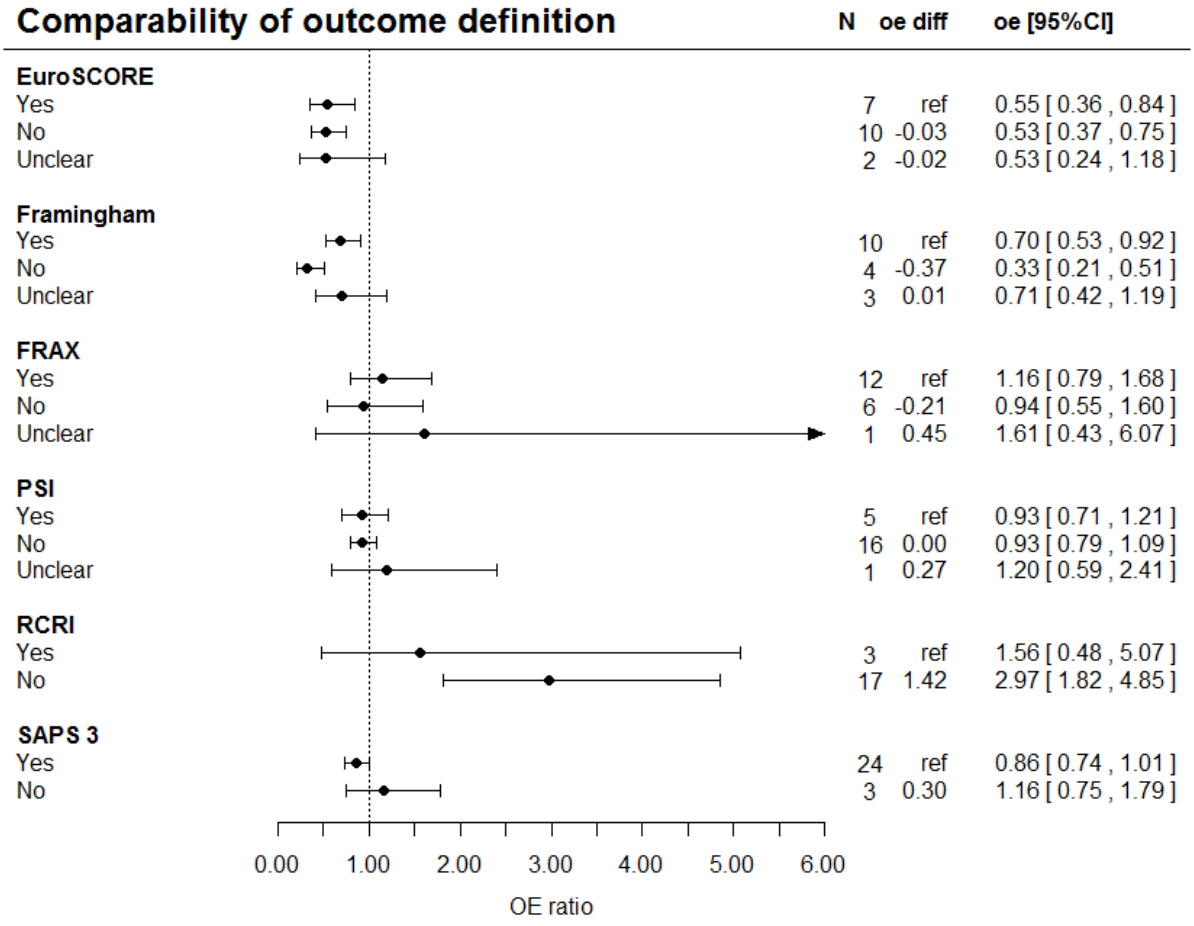

Comparability of age eligibility criteria
N
oe diff
oe [95%CI]
EuroSCORE
Comparable 19 ref 0.54 [ 0.42 , 0.68 ]
Framingham
Narrower 13 ref 0.56 [ 0.41 , 0.76 ]
Mixture 4 0.10 0.66 [ 0.38 , 1.15 ]
FRAX
Narrower 6 ref 1.38 [ 0.82 , 2.35 ]
Comparable 1 0.23 1.61 [ 0.43 , 6.00 ]
Mixture 11 -0.40 0.99 [ 0.67 , 1.46 ]
Broader 1 -0.74 0.65 [ 0.18 , 2.35 ]
PSI
Narrower 2 ref 0.80 [ 0.52 , 1.23 ]
Comparable 14 0.17 0.97 [ 0.82 , 1.14 ]
Broader 6 0.11 0.91 [ 0.70 , 1.18 ]
RCRI
Narrower 1 ref 11.79 [ 1.86 , 74.64 ]
Comparable 3 -10.69 1.10 [ 0.38 , 3.20 ]
Broader 16 -8.88 2.91 [ 1.83 , 4.63 ]
SAPS 3
Narrower 17 ref 0.95 [ 0.79 , 1.15 ]
Comparable 4 -0.15 0.80 [ 0.55 , 1.18 ]
Broader 6 -0.16 0.80 [ 0.58 , 1.10 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio

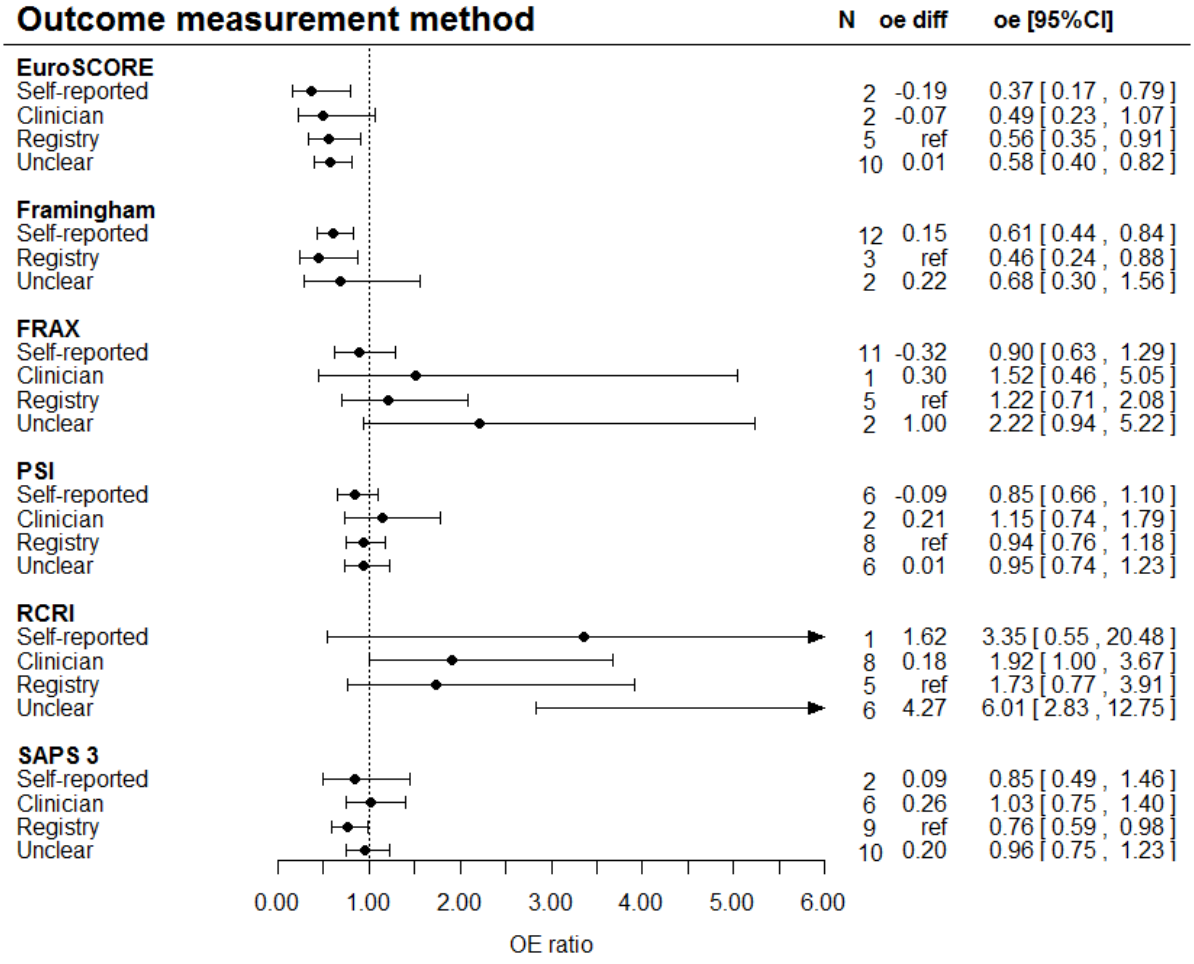


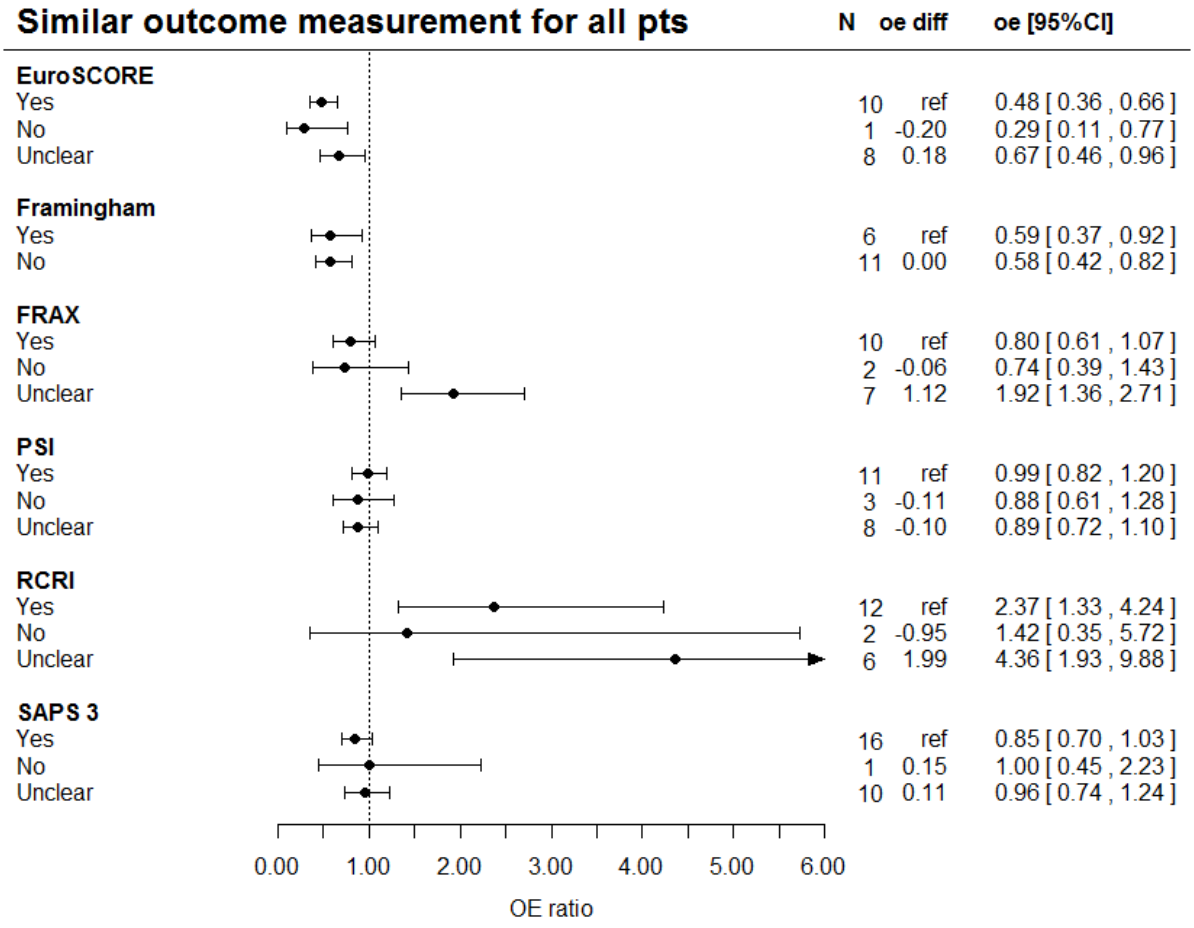

Comparability of continent
N
oe diff
oe [95%CI]
EuroSCORE
Comparable to development
9 ref 0.61 [ 0.43 , 0.86 ]
Non-overlapping
10 -0.13 0.48 [ 0.34 , 0.67 ]
Framingham
Comparable to development
8 ref 0.76 [ 0.54 , 1.07 ]
Non-overlapping
9 -0.29 0.46 [ 0.34 , 0.64 ]
FRAX
Narrower
19 ref 1.10 [ 0.83 , 1.47 ]
PSI
Comparable to development
6 ref 0.99 [ 0.78 , 1.25 ]
Non-overlapping
16 -0.07 0.91 [ 0.78 , 1.07 ]
RCRI
Comparable to development
4 ref 1.79 [ 0.65 , 4.92 ]
Broader
1 -0.54 1.26 [ 0.17 , 9.48 ]
Non-overlapping
15 1.40 3.19 [ 1.88 , 5.44 ]
SAPS 3
Narrower
27 ref 0.89 [ 0.77 , 1.03 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio

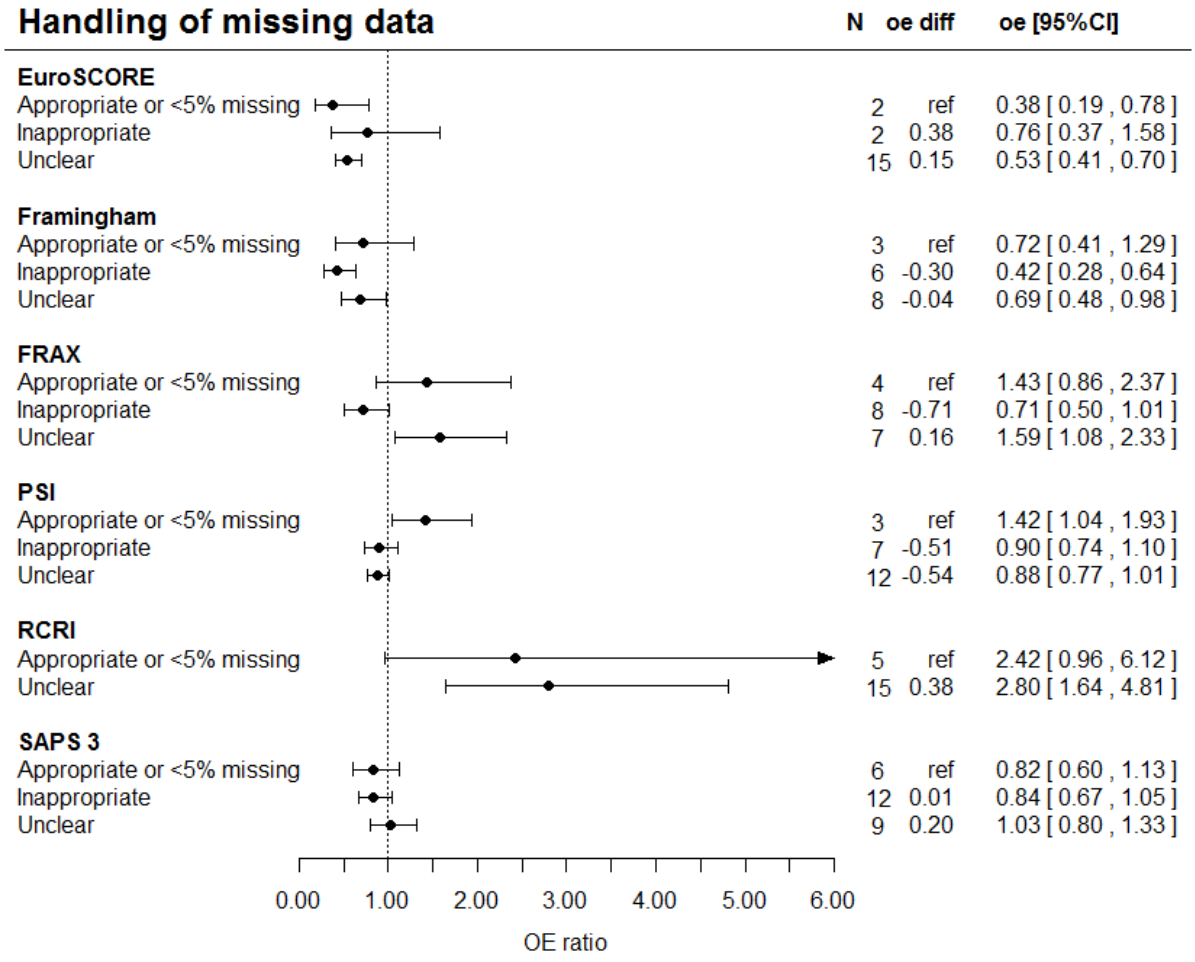



Changes made to predictors
N
oe diff
oe [95%CI]
EuroSCORE
No 11 ref 0.50 [ 0.37 , 0.67 ]
Yes 4 -0.03 0.46 [ 0.28 , 0.76 ]
Unclear 4 0.33 0.82 [ 0.48 , 1.42 ]
Framingham
No 17 ref 0.58 [ 0.45 , 0.76 ]
FRAX
No 14 ref 1.18 [ 0.84 , 1.66 ]
Yes 5 -0.28 0.90 [ 0.51 , 1.59 ]
PSI
No 12 ref 0.93 [ 0.78 , 1.12 ]
Yes 5 -0.04 0.89 [ 0.67 , 1.20 ]
Unclear 5 0.05 0.98 [ 0.75 , 1.29 ]
RCRI
No 16 ref 3.03 [ 1.84 , 4.98 ]
Yes 1 -2.44 0.59 [ 0.08 , 4.10 ]
Unclear 3 -0.52 2.51 [ 0.80 , 7.88 ]
SAPS 3
No 26 ref 0.87 [ 0.76 , 1.00 ]
Unclear 1 1.07 1.94 [ 0.90 , 4.20 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio


Comparability of outcome definition
N
oe diff
oe [95%CI]
EuroSCORE
Yes 7 ref 0.55 [ 0.36 , 0.84 ]
No 10 -0.03 0.53 [ 0.37 , 0.75 ]
Unclear 2 -0.02 0.53 [ 0.24 , 1.18 ]
Framingham
Yes 10 ref 0.70 [ 0.53 , 0.92 ]
No 4 -0.37 0.33 [ 0.21 , 0.51 ]
Unclear 3 0.01 0.71 [ 0.42 , 1.19 ]
FRAX
Yes 12 ref 1.16 [ 0.79 , 1.68 ]
No 6 -0.21 0.94 [ 0.55 , 1.60 ]
Unclear 1 0.45 1.61 [ 0.43 , 6.07 ]
PSI
Yes 5 ref 0.93 [ 0.71 , 1.21 ]
No 16 0.00 0.93 [ 0.79 , 1.09 ]
Unclear 1 0.27 1.20 [ 0.59 , 2.41 ]
RCRI
Yes 3 ref 1.56 [ 0.48 , 5.07 ]
No 17 1.42 2.97 [ 1.82 , 4.85 ]
SAPS 3
Yes 24 ref 0.86 [ 0.74 , 1.01 ]
No 3 0.30 1.16 [ 0.75 , 1.79 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio




Similar outcome measurement for all pts
N
oe diff
oe [95%CI]
EuroSCORE
Yes 10 ref 0.48 [ 0.36 , 0.66 ]
No 1 -0.20 0.29 [ 0.11 , 0.77 ]
Unclear 8 0.18 0.67 [ 0.46 , 0.96 ]
Framingham
Yes 6 ref 0.59 [ 0.37 , 0.92 ]
No 11 0.00 0.58 [ 0.42 , 0.82 ]
FRAX
Yes 10 ref 0.80 [ 0.61 , 1.07 ]
No 2 -0.06 0.74 [ 0.39 , 1.43 ]
Unclear 7 1.12 1.92 [ 1.36 , 2.71 ]
PSI
Yes 11 ref 0.99 [ 0.82 , 1.20 ]
No 3 -0.11 0.88 [ 0.61 , 1.28 ]
Unclear 8 -0.10 0.89 [ 0.72 , 1.10 ]
RCRI
Yes 12 ref 2.37 [ 1.33 , 4.24 ]
No 2 -0.95 1.42 [ 0.35 , 5.72 ]
Unclear 6 1.99 4.36 [ 1.93 , 9.88 ]
SAPS 3
Yes 16 ref 0.85 [ 0.70 , 1.03 ]
No 1 0.15 1.00 [ 0.45 , 2.23 ]
Unclear 10 0.11 0.96 [ 0.74 , 1.24 ]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio




Total OE ratio for categories of study characteristics, pooled using univariable meta-regression analyses per systematic review. N represents the number of external validation studies in a specific category. OE diff represents the difference in OE ratio with regard to a reference category (indicated with 'ref'). Dev: development, val: validation, incr: incremental value, pts: patients.

## References


1. Snell KI, 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? *Stat Methods Med Res* 2017:962280217705678.
2. Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: asymptotic methods and evaluation. *Stat Med* 2006;25(4):559-73.
3. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29-36.
4. 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.
5. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* 2014;14:25.
6. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2015.
7. R: A language and environment for statistical computing [program]. Vienna, Austria: R Foundation for Statistical Computing, 2016.
8. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 2010;36(3):1-48.
9. Gasparrini A, Armstrong B, Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Stat Med* 2012;31(29):3821-39.
10. Debray TP. Metamisc: Diagnostic and Prognostic Meta-Analysis. 2017.
11. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw* 2015;67(1).

| Section/topic | Proposed item to be used in methodology research | Reported on page |
|---|---|---|
| **TITLE** | | |
| Title | Identify the report as a meta-epidemiologic study. | 1 |
| **ABSTRACT** | | |
| Structured summary | Provide a structured summary that includes the background of the topic, goal of the study, data sources, method of data selection, appraisal and synthesis methods, results, limitations, conclusions and implications of key findings. | 3 |
| **INTRODUCTION** | | |
| Rationale | Describe the rationale for the meta-epidemiological study in the context of what is already known. | 5 |
| Objectives | Provide an explicit statement of the goal of the meta-epidemiological study and the hypothesis being empirically tested. | 5 |
| **METHODS** | | |
| Protocol | Indicate if a protocol exists, if and where it can be accessed (eg, Web address). Registration of a protocol is not mandatory. | Available on request |
| Eligibility criteria | Specify study characteristics used as criteria for eligibility with a rationale. | 7 |
| Information sources | Describe all information sources (eg, databases with dates of coverage, contact with experts to identify additional studies, Internet searches) and search date. | 7 |
| Search | Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated. Search is commonly not driven by a clinical question. | Supplement page 2 |
| Study selection | Describe the process for selecting studies for inclusion (ie, how many reviewers selected studies, reviewing in duplicate or by single individuals). | 7,8 |
| Data collection process | Describe method of data extraction from reports (eg, piloted forms, independently, in duplicate) and any processes used for manipulating data or obtaining and confirming data from investigators. | 8,9, Supplement page 3-6 |
| Data items | List and define all variables for which data were sought and any assumptions and imputations made. | Supplement page 3-6 |
| Risk of bias in individual studies | If risk of bias assessment of individual studies was relevant to the analysis, describe the items used and how this information is to be used during data synthesis. | Not assessed |
| Summary measures | State the principal summary measures (eg, ratio of risk ratios, difference in means) and explain its meaning and direction to readers. | 9, Supplement page 7, 8 |
| Synthesis of results | Describe the statistical or descriptive methods of synthesis including measures of consistency if relevant. If applicable, describe the development of statistical or simulation modelling based on theoretical background. Describe and justify assumptions and | 9, Supplement page 7, 8 |

*From: Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. Evid Based Med 2017;22(4):139-42.*

| Section/topic | Proposed item to be used in methodology research | Reported on page |
|---|---|---|
| | computational approximations. Describe methods of additional analyses (eg, sensitivity or subgroup analyses, meta-regression), if done, indicating which were prespecified. | |
| **RESULTS** | | |
| Study selection | Give numbers of studies assessed for eligibility and included in the study, with reasons for exclusions at each stage, ideally with a flow diagram. Present a measure of inter-reviewer agreement (eg, kappa statistic). | 10, Figure 1 |
| Study characteristics | For each study, present characteristics for which data were extracted and provide the citations. Clinical characteristics may not always be relevant. | Supplement page 9-12 |
| Risk of bias within studies | If risk of bias assessment of individual studies was used in the meta-epidemiological analysis, report risk of bias indicators of each study to allow replication of findings. | Not assessed |
| Results of individual studies | Present data elements used in the meta-epidemiological analysis from each study (results of clinical outcomes may not be relevant). | Not done |
| Synthesis of results | Present results of statistical analysis done, including measures of precision and measures of consistency. Present validity of assumptions and fit of statistical or simulation modelling, if applicable. | 11, 12, Figure 2-5, Supplement page 13-49 |
| Additional analysis | Give results of additional analyses, if done (eg, sensitivity or subgroup analyses, meta-regression). | Not done |
| **DISCUSSION** | | |
| Summary of evidence | Summarise the main findings and compare them with existing knowledge about the topic. The quality of evidence may not be relevant; however, investigators should describe their certainty in the results to readers. | 13,14 |
| Limitations | Discuss limitations at research methodology level (eg, likelihood of reporting or publication bias). | 13,14 |
| Conclusions | Provide general interpretation of the results and implications for future research. Provide any plausible impact on clinical practice. | 16 |
| **FUNDING** | | |
| Funding | Describe sources of funding for the methodology research and role of funders. | 17 |

*From: Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. Evid Based Med 2017;22(4):139-42.*