BMJ Open

Novel approach to meta-analysis of tests and clinical prediction rules with three or more risk categories

Mark H Ebell ,[1] Mary E Walsh ,[2,3] Fiona Boland ,[2] Brian McKay,[1] Tom Fahey [2]

To cite: Ebell MH, Walsh ME, Boland F, *et al*. Novel approach to meta-analysis of tests and clinical prediction rules with three or more risk categories. *BMJ Open* 2021;**11**:e036262. doi:10.1136/bmjopen-2019-036262

► Prepublication history and additional material for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2019-036262).

Received 07 December 2019
Revised 09 November 2020
Accepted 02 December 2020

[1]Epidemiology and Biostatistics, University of Georgia, Athens, Georgia, USA
[2]HRB Centre for Primary Care Research, Department of General Practice, Royal College of Surgeons in Ireland, Dublin, Ireland
[3]School of Physiotherapy, Royal College of Surgeons in Ireland, Dublin, Ireland

Correspondence to
Professor Mark H Ebell; ebell@uga.edu

ABSTRACT

Objective Multichotomous tests have three or more outcome or risk categories, and can provide richer information and a better fit with clinical decision-making than dichotomous tests. Our objective is to present a fully developed approach to the meta-analysis of multichotomous clinical prediction rules (CPRs) and tests, including meta-analysis of stratum specific likelihood ratios.

Study design We have developed a novel approach to the meta-analysis of likelihood ratios for multichotomous tests that avoids the need to dichotomise outcome categories, and demonstrate its application to a sample CPR. We also review previously reported approaches to the meta-analysis of the area under the receiver operating characteristic curve (AUROCC) and meta-analysis of a measure of calibration (observed:expected) for multichotomous tests or CPRs.

Results Using data from 10 studies of the Cancer of the Prostate Risk Assessment (CAPRA) risk score for prostate cancer recurrence, we calculated summary estimates of the likelihood ratios for low, moderate and high risk groups of 0.40 (95% CI 0.32 to 0.49), 1.24 (95% CI 0.99 to 1.55) and 4.47 (95% CI 3.21 to 6.23), respectively. Applying the summary estimates of the likelihood ratios for each risk group to the overall prevalence of cancer recurrence in a population allows one to estimate the likelihood of recurrence for each risk group in that population.

Conclusion An approach to meta-analysis of multichotomous tests or CPRs is presented. A spreadsheet for data preparation and code for R and Stata are provided for other researchers to download and use. Combined with summary estimates of the AUROCC and calibration, this is a comprehensive strategy for meta-analysis of multichotomous tests and CPRs.

Strengths and limitations of this study

- ► We present a novel approach to the meta-analysis of stratum specific likelihood ratios for multichotomous tests.
- ► This avoids limitations of previous studies.
- ► It is computationally straightforward and code for R and Stata is provided.

INTRODUCTION

Multichotomous clinical prediction rules (CPRs) and diagnostic tests classify patients into three or more risk categories or risk groups for an outcome. Examples include the Strep score,[1] the Wells score for diagnosis of deep vein thrombosis,[2] the Asymmetry Border Color Diameter (ABCD) rule for the evaluation of skin lesions[3] and the Good Outcome Following Attempted Resuscitation (GO-FAR) score to predict the outcome of in-hospital cardiopulmonary resuscitation.[4] An important advantage of multichotomous test interpretation is that it provides more information than simply dichotomising, and offers greater coherence with recommended strategies for clinical decision-making. The threshold model of decision-making recommends identifying a low risk group in whom disease can be ruled out, a high-risk group in whom it can be ruled in, and an intermediate risk group that requires further testing or information gathering.[5] Multichotomous CPRs with three (or more) risk categories are able to classify patients in a way that reflects these decision thresholds, making them potentially more useful to clinicians.[6]

For example, a CPR was developed to predict the likelihood of being diagnosed with rheumatoid arthritis (RA) 1 year later among patients presenting with undifferentiated joint pain to a general practitioner.[7] Simply dichotomising the risk score into low and high risk groups based on a single cut-off that maximises the sum of sensitivity and specificity creates two risk groups with 11% and 68% probabilities of developing RA. The low risk group is arguably not low risk enough to rule out the diagnosis, and the high-risk group may not be high enough to initiate therapy. Therefore, the authors identified low, moderate and high risk groups (<5, 5 to 9 and >9 points) to identify groups with 3%, 46% and 84% probabilities of subsequent RA. The low risk group now has the disease almost entirely ruled out, patients in the moderate risk group might

be designated for close follow-up and repeat testing, and the high risk group is high enough in risk that one could consider for initiation of a disease modifying anti-rheumatic drug. Thus, the additional information from having more than two outcome categories proves very useful clinically.

While one can calculate positive and negative likelihood ratios for a dichotomous CPR, multichotomous CPRs do not have a single cut-off. Instead, the preferred measure of diagnostic accuracy for multichotomous tests and CPRs is the stratum specific likelihood ratio, that is, the likelihood ratio associated with each risk group. Because likelihood ratios are a characteristic of the test, in theory they should not vary with changes in disease prevalence (and assuming a generally similar spectrum of disease). Previous meta-analyses have taken one or more of the following five approaches to meta-analysis of a multichotomous CPRs, but all have limitations:

1. Calculating the area under a summary receiver operating characteristic (ROC) curve, with each study contributing a single sensitivity/specificity pair to the plot;[8 9]
2. Reporting calibration as a risk ratio (RR), where a RR >1.0 represents overprediction of the diagnosis, and a RR <1.0 underprediction;[10 11]
3. Performing meta-analysis of ROC curves;[12]
4. Dichotomising the test, by combining groups until there are only two dichotomous categories with a single cut-off, and then calculating summary measures of sensitivity, specificity and positive and negative likelihood ratio;[3] and
5. Combining the predictive values of an outcome for a risk group using meta-analysis.[13]

As noted, all of these methods have limitations that affect their interpretability and usefulness. Summary ROC curves are useful for determining discrimination, but do not provide summary estimates of accuracy or calibration. Calibration (the ratio of observed to expected or O:E) is important for evaluating whether a rule is consistent with the performance in the original study, but does not provide an estimate of the likelihood of an outcome for patients in a particular risk group. Meta-analysis of predictive values (the likelihood of disease in a risk group) is inappropriate because predictive values may vary greatly with the underlying prevalence of disease, even if the CPR has the same accuracy as measured by stratum specific likelihood ratios across studies.[13] Finally, dichotomising CPRs that have three or more risk groups into two groups in order to calculate summary estimates of accuracy loses information as noted above, and is inconsistent with how the CPR was intended to be used or interpreted. For example, a clinician might ask: how much does having an ABCD score of 4 points increase the likelihood of melanoma, compared with scores of 2 points or 3 points? If scores of 2, 3 and 4 are combined into a single high risk group to dichotomise the risk score, that information is lost.

In this article, we describe a comprehensive approach to the meta-analysis of multichotomous tests and CPRs. First, we propose a novel approach to the calculation of a summary estimate of the stratum specific likelihood ratio (SSLR) for each risk group of a multichotomous test or CPR. We will also review methods, described in detail by Debray and colleagues,[14 15] for the meta-analysis of the area under the receiver operating characteristic curve (AUROCC) to calculate a summary estimate of discrimination and meta-analysis of the ratio of observed to expected outcomes to calculate a summary estimate of calibration. Finally, we apply our approach to meta-analysis of SSLRs to the Cancer of the Prostate Risk Assessment (CAPRA) score for prostate cancer prognosis.

METHODS

Calculating summary estimates of stratum specific likelihood ratios

A likelihood ratio (LR) is the likelihood of a test result in patients with the disease divided by the likelihood of the test result in patients without the disease.[16] When calculated for a dichotomous test, positive and negative LRs are commonly reported. For a multichotomous test or CPR with more three or more risk categories, each risk category has its own LR, called the SSLR. This section describes development and implementation of a novel approach to the calculation of SSLRs for multichotomous tests.

To calculate summary estimates of the SSLR, we will treat the likelihood ratio as a type of risk ratio, making it possible to adapt methods already developed for meta-analysis of risk ratios in randomised trials. By determining SSLRs, we can then apply them to the overall prevalence of disease in the population and calculate the post-test probability of disease for each risk category using Bayes' formula. It is important to note that when calculating summary estimates of multichotomous (or dichotomous) tests, it is important that the same cut-offs are used across studies. For example, consistently defining low risk as 0 points, moderate risk as 1 to 2 points and high risk as 3 to 4 points. It would be inappropriate to perform meta-analysis when risk groups are defined differently by different studies.

For a dichotomous test, the LR is calculated as follows, where Pr is probability, T+=positive test result, T–=negative test result, D+ is patients with disease and D– is patients without disease (note that 'disease' could represent any outcome predicted by a test or CPR, including death vs survival or treatment benefit vs treatment harm):

$$LR+ = Pr(T+ \mid D+) \,/\, Pr(T+ \mid D-)$$

$$LR- = Pr(T- \mid D+) \,/\, Pr(T- \mid D-)$$

For a multichotomous test or CPR, each risk category has its own SSLR; there is no longer a positive and negative LR. For example, if a CPR places patients into low, moderate and high risk groups, the SSLRs are calculated as follows. Note that $T_{\text{low risk}}$, $T_{\text{moderate risk}}$ and $T_{\text{high risk}}$ are patients classified as low risk, moderate risk or high risk, while D+ is the total number of patients with the outcome and D– is the total without the outcome (for

Table 1 Calculation of stratum specific likelihood ratios for a single study of the Cancer of the Prostate Risk Assessment (CAPRA) score[20] to predict the likelihood that a patient has a biochemical recurrence of prostate cancer

Generic risk group	Recurrence of prostate CA	No recurrence of prostate CA	Stratum specific likelihood ratio
Low risk	a	x	LR_{low} = (a / D+) / (x / D–)
Moderate risk	b	y	LR_{mod} = (b / D+) / (y / D–)
High risk	c	z	LR_{high} = (c / D+) / (z / D–)
	D+	D–	
CAPRA risk group	**Recurrence of prostate CA**	**No recurrence of prostate CA**	**Stratum specific likelihood ratio**
Low (0–2 pts)	69	764	LR_{low} = (69/210)/(764/1229)=0.53
Moderate (3–5 pts)	103	432	LR_{mod} = (103/210)/(432/1229)=1.4
High (6–10 pts)	38	33	LR_{high} = (38/210)/(33/1229)=6.7
	210	1229	

CA, cancer; LR, likelihood ratio; pts, points.

CPRs the outcome being predicted is often the likelihood of disease, hence use of D):

$$LR_{low} = Pr(T_{low\ risk} \mid D+) \ / \ Pr(T_{low\ risk} \mid D-)$$
$$LR_{moderate} = Pr(T_{moderate\ risk} \mid D+) \ / \ Pr(T_{moderate\ risk} \mid D-)$$
$$LR_{high} = Pr(T_{high\ risk} \mid D+) \ / \ Pr(T_{high\ risk} \mid D-)$$

The CAPRA score is a CPR that assigns men with prostate cancer to low (0 to 2 points), moderate (3 to 5 points) or high risk (6 or more points) groups for biochemical recurrence after some period, typically 5 years from the time of initial treatment.[17] Several validation studies of the CAPRA score have been conducted; the calculation of SSLRs for a single study is shown in table 1.[18]

For any multichotomous CPR or test, the SSLR for each risk category is the ratio of two risks or probabilities: for patients in that risk category, the probability of recurrence divided by the probability of no recurrence. This is similar conceptually to a RR for a treatment trial, defined as the ratio of the risk or probability of an outcome in the treatment group to the risk or probability of that outcome in the control group. Table 2 has five parts that illustrate how likelihood ratios can be treated as RRs for the calculation of SSLRs.

Part 1 shows how data are formatted for a meta-analysis of three hypothetical treatment trials with recurrence of prostate cancer as the primary outcome. Part 2 shows the usual approach to displaying results of a study with three or more risk groups, and how the SSLR for a single study are calculated. Part 3 reformats the same data to mimic the RRs of a treatment trial, illustrating how the RRs are identical to the LRs calculated in Part 2. Finally, Part 4 illustrates the general case for formatting the results of a study describing a CPR with three risk categories, and Part 5 illustrates the general form of the equation showing how the same approach can be extended to a test or CPR with any number of risk categories.

A Microsoft Excel spreadsheet that facilitates the preparation of multichotomous data for analysis (in this case 3 risk categories) is available for free download at https://doi.org/10.5281/zenodo.3936001. Column A should be filled in with the study name, Column B with the study year, Column C with the risk group labels, Column D with the number of patients in the risk group with the outcome of interest and Column F with the number of patients in the risk group without the outcome of interest. Columns E, G, H and I are calculated. The 'Optional' Columns J through L can be used to stratify the analysis on an important study variable such as the test's cut-off, age group or reference standard used. Note that as an internal check, the sum of the number of participants in each row should equal the total number of participants in the study as a whole (Column H). Users should create the desired descriptive variable names appropriate for their data in Row 1. The data are now ready to be imported into Stata, SAS or R for analysis.

After importing the data into Stata 15.1 (StataCorp, College Station, Texas) we used the metan procedure (V.9) to perform a random effects meta-analysis of RRs using the following command (a random effects model was chosen as it is more conservative and accounts to some extent for between study as well as within study variance): metan RecurInRiskGroup RecurNotInRiskGroup NoRecurInRiskGroup NoRecurNotInRiskGroup, random by(RiskGroup) sortby(Year) cc(0.5) lcols(AuthorYear) xlabel(0.05, 0.1, 0.2, 0.5, 2.0, 5.0, 10.0)

To create a forest plot for only the low risk stratum, the following command is used: metan RecurInRiskGroup RecurNotInRiskGroup NoRecurInRiskGroup NoRecurNotInRiskGroup if RiskGroup==“Low risk”, random sortby(Year) cc(0.5) lcols(AuthorYear) xlabel(0.05, 0.1, 0.2, 0.5, 2.0, 5.0, 10.0)

For a script to perform these calculations in R, please see the online supplemental appendix.

Meta-analysis of the area under the ROC curve

In 2017, Debray and colleagues published a detailed guide to meta-analysis of prediction model performance.[14] We have previously applied this guide to the meta-analysis of CPRs with more than two risk categories.[19] Measures of discrimination (area under the curve (AUC)) and corresponding measures of uncertainty

Table 2 Developing a method for formatting data from tests or clinical decision rules with three or more outcomes to calculate stratum specific likelihood ratios

Part 1: Calculating risk ratios for a meta-analysis of treatment trials

	Treatment		Control		
Study	**Recurrence**	**No recurrence**	**Recurrence**	**No recurrence**	**Risk ratio calculation**
Study 1	a_1	b_1	c_1	d_1	$RR=(a_1/(a_1+b_1))/(c_1/(c_1d_1))$
Study 2	a_2	b_2	c_2	d_2	$RR=(a_2/(a_2+b_2))/(c_2/(c_2d_2))$
Study 3	a_3	b_3	c_3	d_3	$RR=(a_3/(a_3+b_3))/(c_1/(c_3d_3))$

Part 2: Usual presentation of a test with three or more risk groups to calculate likelihood ratios (as in table 1)

CAPRA risk group	**Recurrence**	**No recurrence**	**Likelihood ratio calculation**
Low	69	764	$LR_{Low}=(69/210)/(764/1229)=0.53$
Moderate	103	432	$LR_{Mod}=(103/210)/(432/1229)=1.4$
High	38	33	$LR_{High}=(38/210)/(33/1229)=6.7$
	210	1229	

Part 3: Alternate presentation of the same data to calculate likelihood ratios, treating them as risk ratios

	Recurrence		**No recurrence**		
CAPRA risk group	**In risk group**	**Not in risk group**	**In risk group**	**Not in risk group**	**Likelihood ratio calculation**
Low	69	141*	764	465*	$LR_{Low}=(69/(69+141))/(764/(764+465))=0.53$
Moderate	103	107†	432	797†	$LR_{Mod}=(103/(103+107))/(432/(432+797))=1.4$
High	38	172‡	33	1196‡	$LR_{High}=(38/(38+172))/(33/(33+1196))=6.7$

Part 4: Generic representation of how to present data for calculation of stratum specific likelihood ratios with three risk groups

	Outcome or diagnosis present		**Outcome or diagnosis absent**		
Risk group	**In risk group**	**Not in risk group**	**In risk group**	**Not in risk group**	**Likelihood ratio calculation**
Risk group 1	$D+_1$	$D+_2 + D+_3$	$D-_1$	$D-_2 + D-_3$	$LR_1 = (D+_1 / (D+_1 + D+_2 + D+_3)) / (D-_1 / (D-_1 + D-_2 + D-_3))$
Risk group 2	$D+_2$	$D+_1 + D+_3$	$D-_2$	$D-_1 + D-_3$	$LR_2 = (D+_2/(D+_1 + D+_2 + D+_3)) / (D-_2 / (D-_1 + D-_2 + D-_3))$
Risk group 3	$D+_3$	$D+_1 + D+_2$	$D-_3$	$D-_1 + D-_2$	$LR_3 = (D+_3 / (D+_1 + D+_2 + D+_3)) / (D-_3 / (D-_1 + D-_2 + D-_3))$

Part 4: Generic representation of how to present data for calculation of stratum specific likelihood ratios with n risk groups

Risk group i	$D+_i$	$\left(\sum_{i=1}^{n} D+_i\right) - D+_i$	$D-_i$	$\left(\sum_{i=1}^{n} D-_i\right) - D-_i$	$LR_i = \frac{\frac{D+_i}{\sum_{i=1}^{n} D+_i}}{\frac{D-_i}{\sum_{i=1}^{n} D-_i}}$

CAPRA = Cancer of the Prostate Risk Assessment
*Sum of number of patients in moderate and high risk groups with recurrence, that is, 103+38=141 for recurrence group.
†Sum of number of patients in low and high risk groups with recurrence, that is, 69+38=107 for recurrence group.
‡Sum of number of patients in low and moderate risk groups with recurrence, that is, 69+103=172 for recurrence group.
LR, likelihood ratio; RR, risk ratio.

(95% CIs or SEs) can be extracted from individual studies, where reported. In order to conduct meta-analysis, AUC values and reported 95% CIs are transformed to the logit scale and the variance of logit AUC calculated. Where measures of uncertainty are not reported, the variance of logit AUC can be estimated using equations proposed by Debray and colleagues.[14] A random effects meta-analysis of logit AUC and variance values is then conducted with restricted maximum likelihood (REML) estimation, which can be completed, for example, using the metaan procedure in Stata 16 (StataCorp, College Station, Texas).[14 20] The pooled logit AUC and 95% CIs are then back-transformed.[14] The proportion of heterogeneity due to between study variation is estimated using the I^2 statistic. This method could be applied to the CAPRA score, which has a time to event outcome, using the updated framework and R code outlined in the 2019 paper by Debray *et al*.[15]

Meta-analysis of calibration between observed and expected outcomes

Calibration of a CPR refers to the level of agreement between predicted probabilities and observed frequencies of the outcome in a validation study. A summary estimate of calibration of a CPR can be calculated through meta-analysis of 'observed: expected ratios'. Our experience, as also highlighted by Debray and colleagues,[14] was that measures of calibration (O:E ratio, calibration slope or plot) are rarely reported in validation studies of CPRs. Most CPR validation studies will only present the observed number of outcomes in a risk group. If the number of outcomes that would have been 'expected' or 'predicted' based on the rule are not reported, they can be derived or estimated using different methods, depending on what information is available from both the derivation and validation studies.

Ideally, a derivation study of a rule with a binary outcome will present the regression coefficient or OR for each predictor in the model and the intercept.[21] In this case, the proportion of participants expected to have the outcome can be calculated by incorporating the mean values of subject characteristics in the prediction model.[14] In the absence of a full model, a derivation study of a rule may report predicted probabilities for each risk stratum, as is reported by Lim and colleagues for the CRB-65 rule.[22] In this case, the expected number of outcomes in each validation study can be calculated by applying the corresponding predicted probability to the numbers of patients in each risk stratum.[11 14] For example, if the derivation study reported 5% risk of the outcome in those in the low risk category, the expected number of outcomes in the low risk category in the validation study is 5% of those in the category.[11]

As recommended by Debray and colleagues,[14] the O:E ratio is calculated for each study on the log scale as follows: log (number of observed outcomes) – log (number of expected outcomes). If not reported, the variance of log (O:E) ratio can be estimated using equations proposed in their guide.[14] A random effects meta-analysis of log O:E and variance values is conducted with REML estimation. We completed this using the metaan procedure in Stata 14, specifying the exponential option to back-transform results to the scale of interest (StataCorp, College Station, Texas).[14 20] Between study heterogeneity is estimated using the I^2 statistic. As poor calibration can occur if the rule is applied in a population with a different baseline risk than the derivation population, meta-analyses of calibration performance can also pre-define subgroups based on factors that could influence this risk.[14] For example, studies that apply the rule in a primary care setting could be meta-analysed separately to those that apply the rule to hospital inpatients. Again, this method could be applied to the CAPRA score, which has a time to event outcome, using the updated framework and R code described in detail by Debray and colleagues.[15] Presentation of results of meta-analysis of AUC and calibration for the CAPRA score is outside of the scope of this paper, where we focus on novel methods of calculating summary estimates for SSLRs.

Patient and public involvement

No patient involved.

RESULTS

Table 3 presents data from 10 validation studies of the CAPRA score, formatted as shown in Parts 3 and 4 of table 2 discussed above. The LRs for low, moderate and high risk groups for prostate cancer recurrence for each study are shown in the final column. Formatted in this fashion, it becomes straightforward to use standard methods for calculating RRs in any statistical package.

The resulting forest plot (figure 1) shows summary estimates of the SSLR for biochemical recurrence of prostate cancer of 0.40 (95% CI 0.32 to 0.49) for the low risk group, 1.24 (95% CI 0.99 to 1.55) for the moderate risk group and 4.47 (95% CI 3.21 to 6.23) for the high-risk group. The I^2 values (84.7%, 96.1% and 90.6% for the low, moderate and high risk groups, respectively) and visual inspection reveal significant heterogeneity, which may reflect differences in the underlying patient populations.

Presentation of results as a forest plot has several strengths. First, it is a familiar format for meta-analysis, allowing a visual assessment of heterogeneity. A formal assessment of heterogeneity is typically provided; for example, in both R and Stata the I^2 statistic is calculated for each stratum and overall. Note that the LRs calculated for the Cooperberg study are identical to those calculated manually in table 2, an internal verification of the accuracy of our approach.[22] A limitation is that the plot is labelled 'Risk Ratio', although this could easily be modified using a graphics programme (development of a native R package is underway).

Table 3 Data for studies of the CAPRA score with the outcome of recurrence-free survival at 5 years, formatted for calculation of stratum specific likelihood ratios using Stata

Author, year	Year	Risk group	Recur in risk group	Recur not in risk group	No recur in risk group	No recur not in risk group	LR
Ishizaki, 2011	2011	Low risk	21	53	64	73	0.61
Ishizaki, 2011	2011	Moderate risk	35	39	71	66	0.91
Ishizaki, 2011	2011	High risk	18	56	2	135	16.7
Loeb, 2010	2010	Low risk	35	71	669	215	0.44
Loeb, 2010	2010	Moderate risk	53	53	197	687	2.2
Loeb, 2010	2010	High risk	18	88	18	866	8.3
Lughezzani, 2010	2010	Low risk	82	419	826	649	0.29
Lughezzani, 2010	2010	Moderate risk	296	205	567	908	1.5
Lughezzani, 2010	2010	High risk	123	378	82	1393	4.4
May, 2007	2007	Low risk	28	379	399	490	0.15
May, 2007	2007	Moderate risk	218	189	409	480	1.2
May, 2007	2007	High risk	161	246	81	808	4.3
Cooperberg, 2006	2006	Low risk	69	141	764	465	0.53
Cooperberg, 2006	2006	Moderate risk	103	107	432	797	1.4
Cooperberg, 2006	2006	High risk	38	172	33	1196	6.7
Zhao, 2007	2007	Low risk	284	580	4449	1424	0.43
Zhao, 2007	2007	Moderate risk	445	419	1329	4544	2.3
Zhao, 2007	2007	High risk	135	729	95	5778	9.7
Halverson, 2011	2011	Low risk	9	86	167	349	0.29
Halverson, 2011	2011	Moderate risk	27	68	240	276	0.61
Halverson, 2011	2011	High risk	59	36	109	407	2.9
Budaus, 2012	2012	Low risk	98	436	1182	1221	0.37
Budaus, 2012	2012	Moderate risk	280	254	990	1413	1.27
Budaus, 2012	2012	High risk	156	378	231	2172	3.0
Krishnan, 2014	2014	Low risk	6	40	45	254	0.87
Krishnan, 2014	2014	Moderate risk	31	15	230	69	0.88
Krishnan, 2014	2014	High risk	9	37	24	275	2.4
Yoshida, 2012	2012	Low risk	19	99	119	266	0.52
Yoshida, 2012	2012	Moderate risk	57	61	200	185	0.93
Yoshida, 2012	2012	High risk	42	76	66	319	2.1

LR, likelihood ratio.

Furthermore, summary estimates of SSLRs can be used to determine the risk of the outcome in a risk category if one knows the overall prevalence of that outcome in the population. In the 10 identified CAPRA validation studies, 17% of men experienced a biochemical recurrence at 5 years. By using the pretest probability of biochemical recurrence of 17% and the SSLRs of 0.40, 1.24 and 4.47, we can use Bayes' formula to calculate the post-test probability of recurrence as 8% in the low risk group, 20% in the moderate risk group and 48% in the high risk group.

DISCUSSION

We have described a comprehensive approach to the meta-analysis of CPRs with more than two risk categories for an outcome. This approach builds on work by others who have developed approaches to calculating summary estimates of calibration (O:E ratio) and discrimination (area under the ROC curve) by adding a novel approach for the calculation of summary estimates of SSLRs.[11 14] It does not require dichotomising data and avoids the inherent problems with meta-analysis of predictive values. While the focus of this article is on meta-analysis of CPRs with three or more risk categories for an outcome, our approach to the calculation of summary estimates of

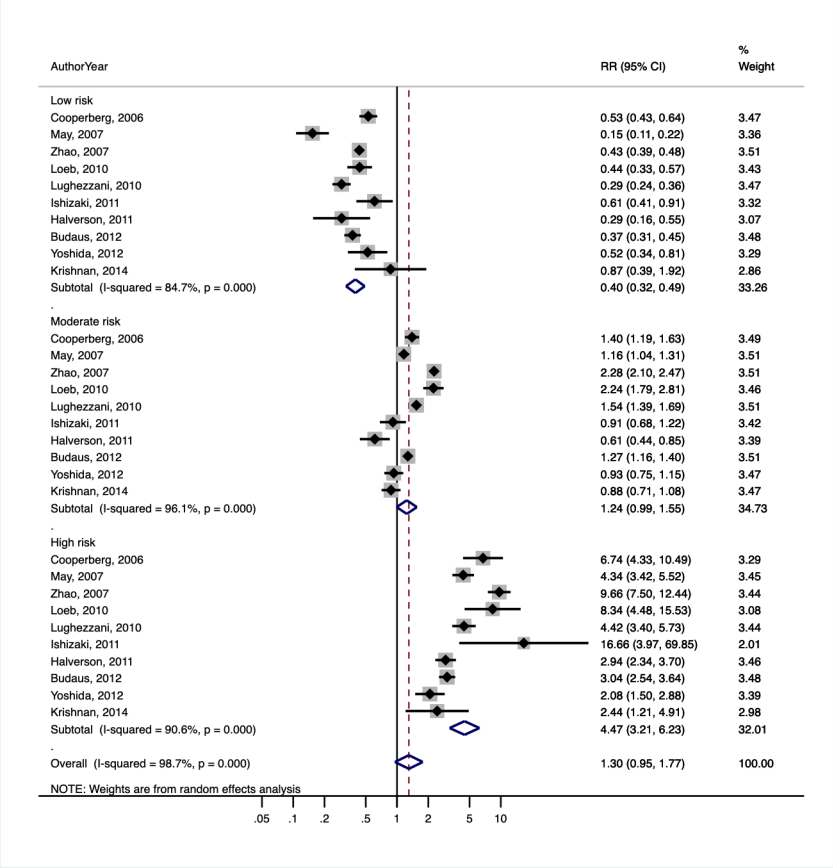

Figure 1 This forest plot shows summary estimates of the stratum specific likelihood ratio for patients classified as low, moderate and high risk for 5-year biochemical recurrence by the CAPRA score. RR,risk ratio.

SSLR could also be applied to any multichotomous diagnostic test such as serum ferritin or d-dimer.[23 24]

Zwinderman and Bossuyt[25] argue that meta-analysis of diagnostic LRs is not appropriate, since the positive and negative LRs are highly correlated for a dichotomous test, because they are calculated from sensitivity and specificity which are also highly correlated. Therefore, they suggest that bivariate meta-analysis of sensitivity and specificity should be performed instead of meta-analysis of LRs, with subsequent calculation of positive and negative LRs if desired. However, this is not relevant for SSLRs that are not calculated from sensitivity or specificity.

Future meta-analyses of multichotomous tests and CPRs should be encouraged to report summary estimates of discrimination, calibration and SSLRs (without dichotomising or collapsing categories) where the underlying data allow these calculations. Each of these metrics provides a different type of information. Discrimination, as measured by a summary estimate of the area under the ROC curve, provides an overall estimate of diagnostic accuracy, and is interpretable for an individual patient by telling us how likely the test or CPR is to correctly classify two randomly selected patients, one with and one without the outcome in question.

Calibration, the agreement between observed and predicted risk, speaks more to how accurately the rule classifies groups of patients with similar levels (for example, deciles) of risk. In some cases, a CPR that has relatively poor discrimination can have excellent calibration. An example is the Breast Cancer Risk Assessment Tool (BCRAT): a meta-analysis found that while the area under the ROC curve is only 0.64, it has very good calibration (O:E 1.08, 95% CI 0.97 to 1.20).[26] Thus, the BCRAT is not helpful when determining the likelihood that an individual woman will be diagnosed with breast cancer in the next 5 years. However, one could state that for 1000 women with a similar BCRAT score, approximately 40 will develop breast cancer in the next 5 years (good calibration), but that we are unable to determine exactly which 40 in this group will develop cancer (poor discrimination).

Furthermore, summary estimates of SSLRs can also be used to determine the likelihood of an outcome in a risk category if one knows the overall prevalence of that outcome in the population. This information is potentially very helpful to clinicians and patients who are trying to interpret the results of a multichotomous test or CPR, and is more easily grasped and applied clinically than concepts such as area under the ROC curve or O:E ratios. And, since the SSLRs are characteristics of the test and are independent of disease prevalence, they can be applied to populations with different prevalences to calculate population-specific post-test probabilities for each risk category.

A limitation of LRs is that while in theory LRs are a feature of the test or risk score and should be consistent across populations (unlike predictive values), in reality it has been shown that there is a degree of variation in LRs between studies.[27] By using a random effects model in our meta-analysis of SSLRs, we do account to some extent for variation. It is also possible to see this variation in the forest plot and see it reflected in the CI of the summary estimate. It is important to note that an important advantage of our approach is that it uses readily available methods in statistical packages to perform the calculations and create the forest plot.

In conclusion, we have developed a novel and easy to use approach to the calculation of summary estimates of SSLRs for any test with three or more outcome categories, and have presented a set of tools that can be applied using standard statistical software to the calculation of summary estimates of SSLRs, discrimination and calibration for multichotomous tests and CPRs.

Acknowledgements The authors would like to acknowledge the contribution of Borislav Dimitrov (deceased) to the development of the methodology for meta-analysis of calibration for multichotomous clinical prediction rules.

Contributors The project was conceptualised and led by Mark Ebell. Brian McKay wrote and tested the R code. Tom Fahey provided input on the conceptualisation, assisted with writing and reviewed the final manuscript. Mary Walsh and Fiona Boland collaborated on the meta-analysis of receiver operating characteristic curves and meta-analysis of observed:expected ratios, and Fiona Boland also helped create Table 2. All co-authors reviewed and approved the final manuscript.

Funding This work was supported in part by a 2019 Fulbright Teaching/Research Scholar award for Dr Ebell (grant number N/A), and funding from the Health Research Board of Ireland to support the HRB Primary Care Research Centre at the Royal College of Surgeons in Ireland, Research Centre Grant no. HRC/2014/1.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no data sets generated and/or analysed for this study. There was no original data collection for this study. R code and an excel spreadsheet for data preparation have been made available to the public under 'Supplemental Files'. The data preparation spreadsheet (Excel) and the R code for stratum specific likelihood ratios can be found at the Zenodo archive: https://doi.org/10.5281/zenodo.3936001.

ORCID iDs
Mark H Ebell http://orcid.org/0000-0003-3228-2877
Mary E Walsh http://orcid.org/0000-0001-8920-7419
Fiona Boland http://orcid.org/0000-0003-3228-0046
Tom Fahey http://orcid.org/0000-0002-5896-5783

REFERENCES

1 Centor RM, Witherspoon JM, Dalton HP, *et al*. The diagnosis of Strep throat in adults in the emergency room. *Med Decis Making* 1981;1:239–46.
2 Wells PS, Anderson DR, Rodger M, *et al*. Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis. *N Engl J Med* 2003;349:1227–35.
3 Harrington E, Clyne B, Wesseling N, *et al*. Diagnosing malignant melanoma in ambulatory care: a systematic review of clinical prediction rules. *BMJ Open* 2017;7:e014096.
4 Ebell MH, Jang W, Shen Y, *et al*. Development and validation of the good outcome following attempted resuscitation (GO-FAR) score to predict neurologically intact survival after in-hospital cardiopulmonary resuscitation. *JAMA Intern Med* 2013;173:1872–8.
5 Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med* 1980;302:1109–17.
6 Ebell M. AHRQ white paper: use of clinical decision rules for point-of-care decision support. *Med Decis Making* 2010;30:712–21.
7 van der Helm-van Mil AHM, le Cessie S, van Dongen H, *et al*. A prediction rule for disease outcome in patients with recent-onset undifferentiated arthritis: how to guide individual treatment decisions. *Arthritis Rheum* 2007;56:433–40.
8 Ebell MH, Culp M, Lastinger K, *et al*. A systematic review of the bimanual examination as a test for ovarian cancer. *Am J Prev Med* 2015;48:350–6.
9 Ebell MH, Culp MB, Radke TJ. A systematic review of symptoms for the diagnosis of ovarian cancer. *Am J Prev Med* 2016;50:384–94.
10 Meurs P, Galvin R, Fanning DM, *et al*. Prognostic value of the CAPRA clinical prediction rule: a systematic review and meta-analysis. *BJU Int* 2013;111:427–36.
11 Dimitrov BD, Motterlini N, Fahey T. A simplified approach to the pooled analysis of calibration of clinical prediction rules for systematic reviews of validation studies. *Clin Epidemiol* 2015;7:267–80.
12 Kester AD, Buntinx F. Meta-Analysis of ROC curves. *Med Decis Making* 2000;20:430–9.
13 van Doorn S, Debray TPA, Kaasenbrood F, *et al*. Predictive performance of the CHA2DS2-VASc rule in atrial fibrillation: a systematic review and meta-analysis. *J Thromb Haemost* 2017;15:1065–77.
14 Debray TPA, Damen JAAG, Snell KIE, *et al*. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;356:i6460.
15 Debray TP, Damen JA, Riley RD, *et al*. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res* 2019;28:2768–86.
16 Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ* 2004;329:168–9.
17 Cooperberg MR, Freedland SJ, Pasta DJ, *et al*. Multiinstitutional validation of the UCSF cancer of the prostate risk assessment for prediction of recurrence after radical prostatectomy. *Cancer* 2006;107:2384–91.
18 Brajtbord JS, Leapman MS, Cooperberg MR. The CAPRA score at 10 years: contemporary perspectives and analysis of supporting studies. *Eur Urol* 2017;71:705–9.
19 Ebell MH, Walsh ME, Fahey T, *et al*. Meta-Analysis of calibration, discrimination, and Stratum-Specific likelihood ratios for the CRB-65 score. *J Gen Intern Med* 2019;34:1304–13.
20 Kontopantelis E, Reeves D. Metaan: Random-effects meta-analysis. *Stata J* 2010;10:395–407.
21 Moons KGM, Altman DG, Reitsma JB, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1–73.
22 Lim WS, van der Eerden MM, Laing R, *et al*. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax* 2003;58:377–82.
23 Kohn MA, Klok FA, van Es N. D-Dimer interval likelihood ratios for pulmonary embolism. *Acad Emerg Med* 2017;24:832–7.
24 Guyatt GH, Oxman AD, Ali M, *et al*. Laboratory diagnosis of iron-deficiency anemia: an overview. *J Gen Intern Med* 1992;7:145–53.
25 Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. *Stat Med* 2008;27:687–97.
26 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;132:365–77.
27 Leeflang MMG, Rutjes AWS, Reitsma JB, *et al*. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ* 2013;185:E537–44.