BMJ Open

Meta-analysis of tests and clinical prediction rules with more than 2 risk categories including a novel approach to meta-analysis of stratum specific likelihood ratios

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-036262
Article Type:	Original research
Date Submitted by the Author:	07-Dec-2019
Complete List of Authors:	Ebell, Mark; University of Georgia, Epidemiology and Biostatistics Walsh, Mary; Royal College of Surgeons in Ireland, HRB Centre for Primary Care Research, Department of General Practice; Royal College of Surgeons in Ireland, School of Physiotherapy Boland, Fiona; Royal College of Surgeons Ireland, 123 St Stephens Green, HRB Centre For Primary Care Research, Division of Population Health Sciences (PHS) Fahey, Tom; Royal College of Surgeons in Ireland, Department of General Practice
Keywords:	PRIMARY CARE, STATISTICS & RESEARCH METHODS, Urological tumours < ONCOLOGY

Meta-analysis of tests and clinical prediction rules with more than 2 risk categories including a novel approach to meta-analysis of stratum specific likelihood ratios

Mark H. Ebell (1)

Mary E. Walsh (2)

Fiona Boland (2)

Tom Fahey (2)

1. Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia, Athens, GA, USA. 2. HRB Centre for Primary Care Research, Royal College of Surgeons in Ireland, 123 St. Stephens Green, Dublin 2, Ireland

Corresponding Author:

Mark H. Ebell MD, MS

125 B.S. Miller Hall

UGA Health Sciences Campus

Athens, GA 30602

706-247-4953

ebell@uga.edu

Word count: 3323 words (not including abstract or references), 3 tables and 1 figure

Keywords: meta-analysis, likelihood ratio, diagnostic accuracy, systematic review, multichotomous

Abstract

Objective

Multichotomous tests have 3 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 (SSLRs).

Study design

We have developed a novel approach to the meta-analysis of SSLRs for multichotomous tests that avoids the need to dichotomize 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 CAPRA risk score for prostate cancer recurrence, we calculated summary estimates of the SSLRs for low, moderate and high-risk groups of 0.40, 1.24 and 4.47 respectively. Applying the SSLRs 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

- 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 3 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 ABCD rule for the evaluation of skin lesions, [3] and the 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 dichotomizing, 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) one year later among patients presenting with undifferentiated joint pain to a general practitioner.[7] Simply dichotomizing the risk score into low and high risk groups based on a single cutoff that maximizes 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 2 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 cutoff. Instead, the preferred measure of diagnostic accuracy for multichotomous tests and CPRs is the stratum specific likelihood ratio, i.e. the likelihood ratio associated with each risk group. Because likelihood ratios are a characteristic of the test, they do not vary with changes in disease prevalence. 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 over-prediction of the diagnosis, and a RR < 1.0 under-prediction [10, 11];
3) performing meta-analysis of receiver operating characteristic curves [12],
4) dichotomizing the test, by combining groups until there are only two dichotomous categories with a single cutoff, 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 (3). Finally, dichotomizing CPRs that have 3 or more risk groups into 2 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 dichotomize 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 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 CAPRA score for prostate cancer prognosis.

Methods

Calculating Summary Estimates of Stratum Specific Likelihood Ratios (SSLR)

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.[19] When calculated for a dichotomous test, positive and negative likelihood ratios are commonly reported. For a multichotomous test or CPR with more 3 or more risk categories, each risk category has its own likelihood ratio, called the “stratum specific likelihood ratio” (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 SSRL, we will treat the diagnostic likelihood ratio as a type of risk ratio, making it possible to adapt methods already developed for meta-analysis of risk ratios in randomized 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.

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 likelihood ratio. 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 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 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, moderate, or high risk groups for biochemical recurrence after some period, typically 5 years from the time of initial treatment.[20] Several validation studies of the CAPRA score have been conducted; the calculation of SSLRs for a single study is shown in Table 1.[21]

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 risk ratio (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 risk ratios for the calculation of SSLRs.

Part 1 shows how data are formatted for a meta-analysis of 3 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 3 or more risk groups, and how the stratum specific likelihood ratios for a single study are calculated. Part 3 reformats the same data to mimic the risk ratios of a treatment trial, illustrating how the risk ratios are identical to the likelihood ratios calculated in Part 2. Finally, Part 4 illustrates the general case for formatting the results of a study describing a CPR with 3 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 http://____. 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 cutoff, 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 TX) we used the metan procedure (version 9) to perform a random effects meta-analysis of risk ratios using the following command:

```
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 Appendix. SAS has no intrinsic features for meta-analysis. Prof. Stephen Senn and colleagues produced a suite of detailed macros, which can be downloaded from: http://www.senns.demon.co.uk/SAS%20Macros/SASMacros.html.

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.[15] Measures of discrimination (AUC) and corresponding measures of uncertainty (95% confidence intervals or standard errors) can be extracted from individual studies, where reported. In order to conduct meta-analysis, AUC values and reported 95% confidence intervals 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 REML estimation, which can be completed using the metaan procedure in Stata 14 (Stata Corp, College Station TX).[14, 16] The pooled logit AUC and 95% confidence intervals are then back-transformed [14]. The proportion of heterogeneity due to between study variation is estimated using the I^2 statistic.

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 (observed:

expected [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 odds ratio for each predictor in the model and the intercept.[17] 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. [18] 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 (Stata Corp, College Station TX).[14, 16] 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.

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 likelihood ratios 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 risk ratios 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 0.32-0.49) for the low risk group, 1.24 (95% CI 0.99-1.55) for the moderate risk group, and 4.47 (95% CI 3.21-6.23) 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 also provided, as I^2 is calculated for each stratum and overall. Note that the likelihood ratios calculated for the Cooperberg study [20] are identical to those calculated manually in Table 2, an internal verification of the accuracy of our approach. A limitation is that the plot is labeled “Risk Ratio”, although this could easily be modified using a graphics program (development of a native R package is underway).

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 2 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) [11] and discrimination (area under the ROC curve) [14] by adding a novel approach for the calculation of summary estimates of stratum specific likelihood ratios. It does not require dichotomizing 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 3 or more risk categories for an outcome, our approach to the calculation of summary estimates of SSLR could also be applied to any multichotomous diagnostic test such as serum ferritin or d-dimer.[22, 23]

Future meta-analyses of multichotomous tests and CPRs should be encouraged to report summary estimates of discrimination, calibration, and stratum specific likelihood ratios (without dichotomizing 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 0.97-1.20). [24] 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.

In conclusion we have developed a novel approach to the calculation of summary estimates of stratum specific likelihood ratios for any test with 3 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.

Acknowledgement

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 CPRs.

Financial support

This work was supported in part by a 2019 Fulbright Teaching/Research Scholar award for Dr. Ebell, 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.

Table 1. Calculation of stratum specific likelihood ratios for a single study of the 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	

Table 2. Developing a method for formatting data from tests or clinical decision rules with 3 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	$RR = [a_1/(a_1+b_1)]/[c_1/(c_1+d_1)]$
Study 2	a_2	b_2	c_2	d_2	$RR = [a_2/(a_2+b_2)]/[c_2/(c_2+d_2)]$
Study 3	a_3	b_3	c_3	d_3	$RR = [a_3/(a_3+b_3)]/[c_3/(c_3+d_3)]$
Part 2. Usual presentation of a test with 3 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 3 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+$_i$	$(\sum_{i=1}^{n} D+_i) - D+_i$	D-$_i$	$(\sum_{i=1}^{n} D-_i) - D-_i$	$LR_i = \frac{D+_i / \sum_{i=1}^{n} D+_i}{D-_i / \sum_{i=1}^{n} D-_i}$

* Sum of number of patients in moderate and high-risk groups with recurrence, i.e. 103 + 38 = 141 for recurrence group

** Sum of number of patients in low and high-risk groups with recurrence, i.e. 69 + 38 = 107 for recurrence group

+ Sum of number of patients in low and moderate-risk groups with recurrence, i.e. 69 + 103 = 172 for recurrence group

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.

AuthorYear	Year	RiskGroup	RecurInRiskGroup	RecurNotInRiskGroup	NoRecurInRiskGroup	NoRecurNotInRiskGroup	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

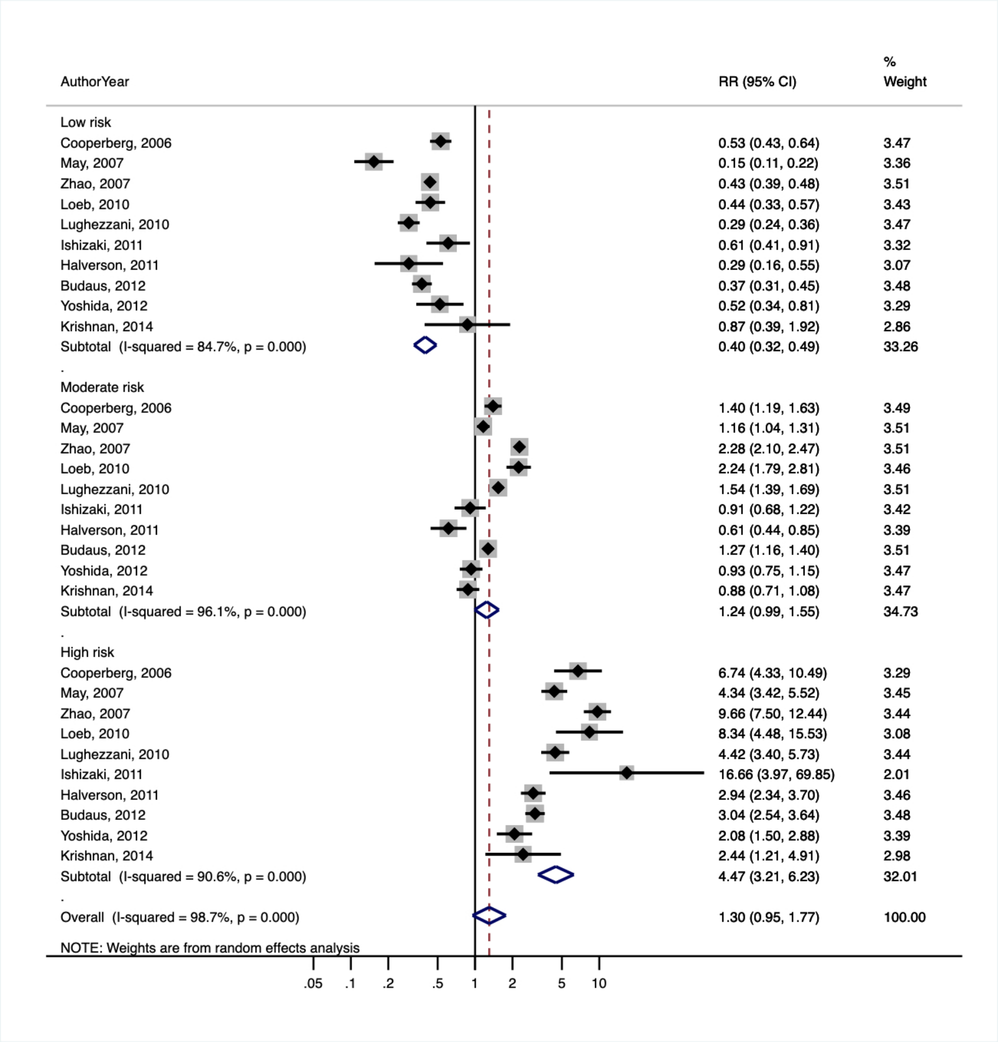
Figure 1 legend

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.

References

1. Centor RM, Witherspoon JM, Dalton HP, Brody CE, Link K: **The diagnosis of strep throat in adults in the emergency room**. *Medical decision making : an international journal of the Society for Medical Decision Making* 1981, **1**(3):239-246.
2. Wells PS, Anderson DR, Rodger M, Forgie M, Kearon C, Dreyer J, Kovacs G, Mitchell M, Lewandowski B, Kovacs MJ: **Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis**. *The New England journal of medicine* 2003, **349**(13):1227-1235.
3. Harrington E, Clyne B, Wesseling N, Sandhu H, Armstrong L, Bennett H, Fahey T: **Diagnosing malignant melanoma in ambulatory care: a systematic review of clinical prediction rules**. *BMJ open* 2017, **7**(3):e014096.
4. Ebell MH, Jang W, Shen Y, Geocadin RG, Get With the Guidelines-Resuscitation I: **Development and validation of the Good Outcome Following Attempted Resuscitation (GO-FAR) score to predict neurologically intact survival after in-hospital cardiopulmonary resuscitation**. *JAMA internal medicine* 2013, **173**(20):1872-1878.
5. Pauker SG, Kassirer JP: **The threshold approach to clinical decision making**. *The New England journal of medicine* 1980, **302**(20):1109-1117.
6. Ebell M: **AHRQ White Paper: Use of clinical decision rules for point-of-care decision support**. *Medical decision making : an international journal of the Society for Medical Decision Making* 2010, **30**(6):712-721.
7. van der Helm-van Mil AH, le Cessie S, van Dongen H, Breedveld FC, Toes RE, Huizinga TW: **A prediction rule for disease outcome in patients with recent-onset undifferentiated arthritis: how to guide individual treatment decisions**. *Arthritis and rheumatism* 2007, **56**(2):433-440.
8. Ebell MH, Culp M, Lastinger K, Dasigi T: **A systematic review of the bimanual examination as a test for ovarian cancer**. *American journal of preventive medicine* 2015, **48**(3):350-356.
9. Ebell MH, Culp MB, Radke TJ: **A Systematic Review of Symptoms for the Diagnosis of Ovarian Cancer**. *American journal of preventive medicine* 2016, **50**(3):384-394.
10. Meurs P, Galvin R, Fanning DM, Fahey T: **Prognostic value of the CAPRA clinical prediction rule: a systematic review and meta-analysis**. *BJU international* 2013, **111**(3):427-436.
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**:267-280.
12. Kester AD, Buntinx F: **Meta-analysis of ROC curves**. *Medical decision making : an international journal of the Society for Medical Decision Making* 2000, **20**(4):430-439.
13. van Doorn S, Debray TPA, Kaasenbrood F, Hoes AW, Rutten FH, Moons KGM, Geersing GJ: **Predictive performance of the CHA2DS2-VASc rule in atrial fibrillation: a systematic review and meta-analysis**. *Journal of thrombosis and haemostasis : JTH* 2017, **15**(6):1065-1077.

14. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, Riley RD, Moons KG: **A guide to systematic review and meta-analysis of prediction model performance**. *Bmj* 2017, **356**:i6460.
15. Ebell MH WM, Fahey T, Kearney M, Marchello C: **Meta-analysis of Calibration, Discrimination, and Stratum-Specific Likelihood Ratios for the CRB-65 Score**. *Annals of family medicine* 2019((in press)).
16. Kontopantelis EaR, D.: **metaan: Random-effects meta-analysis**. *The Stata Journal* 2010, **10**(3):395-407.
17. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS: **Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration**. *Annals of internal medicine* 2015, **162**(1):W1-73.
18. Lim WS, van der Eerden MM, Laing R, Boersma WG, Karalus N, Town GI, Lewis SA, Macfarlane JT: **Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study**. *Thorax* 2003, **58**(5):377-382.
19. Deeks JJ, Altman DG: **Diagnostic tests 4: likelihood ratios**. *Bmj* 2004, **329**(7458):168-169.
20. Cooperberg MR, Freedland SJ, Pasta DJ, Elkin EP, Presti JC, Jr., Amling CL, Terris MK, Aronson WJ, Kane CJ, Carroll PR: **Multiinstitutional validation of the UCSF cancer of the prostate risk assessment for prediction of recurrence after radical prostatectomy**. *Cancer* 2006, **107**(10):2384-2391.
21. Brajtbord JS, Leapman MS, Cooperberg MR: **The CAPRA Score at 10 Years: Contemporary Perspectives and Analysis of Supporting Studies**. *European urology* 2017, **71**(5):705-709.
22. Kohn MA, Klok FA, van Es N: **D-dimer Interval Likelihood Ratios for Pulmonary Embolism**. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine* 2017, **24**(7):832-837.
23. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C: **Laboratory diagnosis of iron-deficiency anemia: an overview**. *Journal of general internal medicine* 1992, **7**(2):145-153.
24. Meads C, Ahmed I, Riley RD: **A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance**. *Breast cancer research and treatment* 2012, **132**(2):365-377.



AuthorYear	RR (95% CI)	% Weight
Low risk		
Cooperberg, 2006	0.53 (0.43, 0.64)	3.47
May, 2007	0.15 (0.11, 0.22)	3.36
Zhao, 2007	0.43 (0.39, 0.48)	3.51
Loeb, 2010	0.44 (0.33, 0.57)	3.43
Lughezzani, 2010	0.29 (0.24, 0.36)	3.47
Ishizaki, 2011	0.61 (0.41, 0.91)	3.32
Halverson, 2011	0.29 (0.16, 0.55)	3.07
Budaus, 2012	0.37 (0.31, 0.45)	3.48
Yoshida, 2012	0.52 (0.34, 0.81)	3.29
Krishnan, 2014	0.87 (0.39, 1.92)	2.86
Subtotal (I-squared = 84.7%, p = 0.000)	0.40 (0.32, 0.49)	33.26
.		
Moderate risk		
Cooperberg, 2006	1.40 (1.19, 1.63)	3.49
May, 2007	1.16 (1.04, 1.31)	3.51
Zhao, 2007	2.28 (2.10, 2.47)	3.51
Loeb, 2010	2.24 (1.79, 2.81)	3.46
Lughezzani, 2010	1.54 (1.39, 1.69)	3.51
Ishizaki, 2011	0.91 (0.68, 1.22)	3.42
Halverson, 2011	0.61 (0.44, 0.85)	3.39
Budaus, 2012	1.27 (1.16, 1.40)	3.51
Yoshida, 2012	0.93 (0.75, 1.15)	3.47
Krishnan, 2014	0.88 (0.71, 1.08)	3.47
Subtotal (I-squared = 96.1%, p = 0.000)	1.24 (0.99, 1.55)	34.73
.		
High risk		
Cooperberg, 2006	6.74 (4.33, 10.49)	3.29
May, 2007	4.34 (3.42, 5.52)	3.45
Zhao, 2007	9.66 (7.50, 12.44)	3.44
Loeb, 2010	8.34 (4.48, 15.53)	3.08
Lughezzani, 2010	4.42 (3.40, 5.73)	3.44
Ishizaki, 2011	16.66 (3.97, 69.85)	2.01
Halverson, 2011	2.94 (2.34, 3.70)	3.46
Budaus, 2012	3.04 (2.54, 3.64)	3.48
Yoshida, 2012	2.08 (1.50, 2.88)	3.39
Krishnan, 2014	2.44 (1.21, 4.91)	2.98
Subtotal (I-squared = 90.6%, p = 0.000)	4.47 (3.21, 6.23)	32.01
.		
Overall (I-squared = 98.7%, p = 0.000)	1.30 (0.95, 1.77)	100.00

NOTE: Weights are from random effects analysis

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.

229x239mm (144 x 144 DPI)

Appendix A. Calculating stratum specific likelihood ratios using R

For analysis in R, we loaded the meta library, created a data frame, and then subsets for each risk group. The metabin procedure was used to calculate summary estimates of stratum specific likelihood ratios, again treating them as risk ratios. Note that the variables are passed to the metabin function as events and total rather than events and non-events as in Stata.

```
library(meta)
# Create data frame and subset for each risk group
capra.df <- data.frame(CAPRA)
capra_low <- subset(capra.df, RiskGroup=="Low risk")
capra_mod <- subset(capra.df, RiskGroup=="Moderate risk")
capra_high <- subset(capra.df, RiskGroup=="High risk")
# Low risk group
meta_low <- metabin(NoRecurInRiskGroup, NoRecurInRiskGroup +
       NoRecurNotInRiskGroup, RecurInRiskGroup, RecurInRiskGroup +
       RecurNotInRiskGroup,
            data = capra_low,
            method = "Inverse")
summary(meta_low)
forest(meta_low)
# Moderate risk group
meta_mod <- metabin(NoRecurInRiskGroup, NoRecurInRiskGroup +
       NoRecurNotInRiskGroup, RecurInRiskGroup, RecurInRiskGroup +
       RecurNotInRiskGroup,
              data = capra_mod,
              method = "Inverse")
summary(meta_mod)
forest(meta_mod)
# High risk group
meta_high <- metabin(NoRecurInRiskGroup, NoRecurInRiskGroup +
       NoRecurNotInRiskGroup, RecurInRiskGroup, RecurInRiskGroup +
       RecurNotInRiskGroup,
              data = capra_high,
              method = "Inverse")
summary(meta_high)
```

forest(meta_high)

Column

A

B

C

E

Optional

D, F - K

Note

Note

Instruction

Enter the study name,

Label describing risk group

Enter the number of patients in each risk group with the outcome of interest, and

Enter the number of patients in each risk group without the outcome of interest.

The “Optional” columns can be used to stratify the analysis on an important study variable, such as duration of follow-up, age group, or test.

These fields are calculated, do not change the formulas

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.

Users should create the desired descriptive variable names appropriate for their data in Row 1.

AuthorYear	Year	RiskGroup	RecurInRiskGroup	RecurNotInRiskGroup
Ishizaki, 2011	2011	Low risk	21	53
Ishizaki, 2011	2011	Moderate risk	35	39
Ishizaki, 2011	2011	High risk	18	56
		Low risk		0
		Moderate risk		0
		High risk		0
		Low risk		0
		Moderate risk		0
		High risk		0
		Low risk		0
		Moderate risk		0
		High risk		0
		Low risk		0
		Moderate risk		0
		High risk		0
		Low risk		0
		Moderate risk		0
		High risk		0
		Low risk		0
		Moderate risk		0
		High risk		0
		Low risk		0
		Moderate risk		0
		High risk		0
		Low risk		0
		Moderate risk		0
		High risk		0
		Low risk		0
		Moderate risk		0
		High risk		0
		Low risk		0
		Moderate risk		0
		High risk		0
		Low risk		0
		Moderate risk		0
		High risk		0
		Low risk		0
		Moderate risk		0
		High risk		0
		Low risk		0
		Moderate risk		0
		High risk		0
		Low risk		0
		Moderate risk		0
		High risk		0

Low risk	0
Moderate risk	0
High risk	0

NoRecurInRIskGroup	NoRecurNotInRiskGroup	Total	LR	Optional1
64	73	211	0.61	
71	66	211	0.91	
2	135	211	16.66	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	
	0	0	#DIV/0!	

0	0	#DIV/0!
0	0	#DIV/0!
0	0	#DIV/0!

Optional2 **Optional3**

BMJ Open

A Novel Approach to Meta-Analysis of Tests and Clinical Prediction Rules with 3 or More Risk Categories

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-036262.R1
Article Type:	Original research
Date Submitted by the Author:	29-Jul-2020
Complete List of Authors:	Ebell, Mark; University of Georgia, Epidemiology and Biostatistics Walsh, Mary; Royal College of Surgeons in Ireland, HRB Centre for Primary Care Research, Department of General Practice; Royal College of Surgeons in Ireland, School of Physiotherapy Boland, Fiona; Royal College of Surgeons in Ireland, HRB Centre For Primary Care Research, Division of Population Health Sciences (PHS) McKay, Brian; University of Georgia, Epidemiology and Biostatistics Fahey, Tom; Royal College of Surgeons in Ireland, Department of General Practice
<b>Primary Subject Heading</b>:	Evidence based practice
Secondary Subject Heading:	Diagnostics, General practice / Family practice, Health services research, Oncology
Keywords:	PRIMARY CARE, STATISTICS & RESEARCH METHODS, Urological tumours < ONCOLOGY

A Novel Approach to Meta-Analysis of Tests and Clinical Prediction Rules With 3 or More Risk Categories

Mark H. Ebell (1)

Mary E. Walsh (2)

Fiona Boland (2)

Brian McKay (1)

Tom Fahey (2)

1. Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia, Athens, GA, USA.

2. HRB Centre for Primary Care Research, Royal College of Surgeons in Ireland, 123 St. Stephens Green, Dublin 2, Ireland

Corresponding Author:

Mark H. Ebell MD, MS
125 B.S. Miller Hall
UGA Health Sciences Campus
Athens, GA 30602
706-247-4953
ebell@uga.edu

Word count: 3623 words (not including abstract or references), 3 tables and 1 figure

Keywords: meta-analysis, likelihood ratio, diagnostic accuracy, systematic review, multichotomous

Standfirst

Diagnostic tests and clinical prediction rules are increasingly presenting multichotomous results, for example low, moderate and high-risk groups. We propose a novel approach to the meta-analysis of stratum specific likelihood ratios and provide code to implement this in R and Stata.

Key Messages

- Diagnostic tests and clinical prediction rules that present the results of a test with more than 2 possible outcomes (multichotomous tests) often report stratum specific likelihood ratios.
- Meta-analysis of stratum specific likelihood ratios for multichotomous tests has usually involved collapsing categories to make the test dichotomous, which loses information.
- We propose a simple approach to calculating stratum specific likelihood ratios for multichotomous tests by formatting data in a way that one can utilize existing software for meta-analysis of risk ratios.
- Methods for meta-analysis of observed to expected ratios and receiver operating curves for multichotomous tests are also reviewed.
- A formatting spreadsheet and code for R and Stata are provided for calculating stratum specific likelihood ratios.

Introduction

Multichotomous clinical prediction rules (CPRs) and diagnostic tests classify patients into 3 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 ABCD rule for the evaluation of skin lesions, [3] and the 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 dichotomizing, 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) one year later among patients presenting with undifferentiated joint pain to a general practitioner.[7] Simply dichotomizing the risk score into low and high risk groups based on a single cutoff that maximizes 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 2 outcome categories proves very useful clinically.

Limitations of previous approaches to meta-analysis of stratum specific likelihood ratios

While one can calculate positive and negative likelihood ratios for a dichotomous CPR, multichotomous CPRs do not have a single cutoff. Instead, the preferred measure of diagnostic accuracy for multichotomous tests and CPRs is the stratum specific likelihood ratio, i.e. 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 over-prediction of the diagnosis, and a RR < 1.0 under-prediction [10, 11];
3) performing meta-analysis of receiver operating characteristic curves [12],
4) dichotomizing the test, by combining groups until there are only two dichotomous categories with a single cutoff, 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 (3). Finally, dichotomizing CPRs that have 3 or more risk groups into 2 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 dichotomize the risk score, that information is lost.

Overview

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 then apply our approach to the CAPRA score for prostate cancer prognosis. We will also review methods for the meta-analysis of the area under the receiver operating characteristic curve (AUROCC) to calculate a summary estimate of discrimination, as well as meta-analysis of the ratio of observed to expected outcomes to calculate a summary estimate of calibration.

Calculating Summary Estimates of Stratum Specific Likelihood Ratios (SSLR)

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.[19] When calculated for a dichotomous test, positive and negative likelihood ratios are commonly reported. For a multichotomous test or CPR with more 3 or more risk categories, each risk category has its own likelihood ratio, called the “stratum specific likelihood ratio” (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 SSRL, 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 randomized 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 cutoffs 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 likelihood ratio. 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_{high}

$_{\text{risk}}$ are patients classified 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 CPRs the outcome being predicted is often the likelihood of disease, hence use of D):

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

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

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

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 risk ratio (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 risk ratios for the calculation of SSLRs.

Part 1 shows how data are formatted for a meta-analysis of 3 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 3 or more risk groups, and how the stratum specific likelihood ratios for a single study are calculated. Part 3 reformats the same data to mimic the risk ratios of a treatment trial, illustrating how the risk ratios are identical to the likelihood ratios calculated in Part 2. Finally, Part 4 illustrates the general case for formatting the results of a study describing a CPR with 3 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 (see Supplemental Files below). 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 cutoff, 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 TX) we used the metan procedure (version 9) to perform a random effects meta-analysis of risk ratios using the following command (a random effects model was chosen as it is more conservative and does not accounts to some extent for between study variability):

```
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 Appendix. SAS has no intrinsic features for meta-analysis.

Results of Application to the CAPRA score

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.[20] Several validation studies of the CAPRA score have been conducted.[21]

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 likelihood ratios 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 risk ratios 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 0.32-0.49) for the low risk group, 1.24 (95% CI 0.99-1.55) for the moderate risk group, and 4.47 (95% CI 3.21-6.23) 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 likelihood ratios calculated for the Cooperberg study [20] are identical to those calculated manually in Table 2, an internal verification of the accuracy of our approach. A limitation is that the plot is labeled “Risk Ratio”, although this could easily be modified using a graphics program (development of a native R package is underway).

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.

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.[15] Measures of discrimination (AUC) and corresponding measures of uncertainty (95% confidence intervals or standard errors) can be extracted from individual studies, where reported. In order to conduct meta-analysis, AUC values and reported 95% confidence intervals 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 REML estimation, which can be completed for example using the metaan procedure in Stata 16 (Stata Corp, College Station

TX).[14, 16] The pooled logit AUC and 95% confidence intervals 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.[26]

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 (observed: expected [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 odds ratio for each predictor in the model and the intercept.[17] 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. [18] 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 (Stata Corp, College Station TX).[14, 16] 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 [26]. Presentation of results of meta-analysis of area under the curve 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.

Discussion

We have described a comprehensive approach to the meta-analysis of CPRs with more than 2 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) [11] and discrimination (area under the ROC curve) [14] by adding a novel approach for the calculation of summary estimates of stratum specific likelihood ratios. It does not require dichotomizing 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 3 or more risk categories for an outcome, our approach to the calculation of summary estimates of SSLR could also be applied to any multichotomous diagnostic test such as serum ferritin or d-dimer.[22, 23]

Zwinderman and Bossuyt argue that meta-analysis of diagnostic likelihood ratios is not appropriate, since the positive and negative likelihood ratios 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 likelihood ratios, with subsequent calculation of positive and negative likelihood ratios if desired. However, this is not relevant for stratum specific likelihood ratios that are not calculated from sensitivity or specificity.[25].

Future meta-analyses of multichotomous tests and CPRs should be encouraged to report summary estimates of discrimination, calibration, and stratum specific likelihood ratios (without dichotomizing 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-

1.20). [24] 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.

In conclusion we have developed a novel approach to the calculation of summary estimates of stratum specific likelihood ratios for any test with 3 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.

Supplemental materials

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

Contributorship Statement

The project was conceptualized and led by Mark Ebell. Brian McKay wrote and tested the R code. Tom Fahey provided input on the conceptualization, assisted with writing, and reviewed the final manuscript. Mary Walsh and Fiona Boland collaborated on the meta-analysis of ROC 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.

Acknowledgement

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 CPRs.

Patient and Public Involvement

No patient involved.

Funding

This work was supported in part by a 2019 Fulbright Teaching/Research Scholar award for Dr. Ebell, 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.

Data sharing agreement

There was no original data collection for this study. R code and anexcel spreadsheet for data preparation have been made available to the public under “Supplemental Files”.

Table 1. Calculation of stratum specific likelihood ratios for a single study of the 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	

Table 2. Developing a method for formatting data from tests or clinical decision rules with 3 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 = $[a_1/(a_1+b_1)]/[c_1/(c_1+d_1)]$
Study 2	a_2	b_2	c_2	d_2	RR = $[a_2/(a_2+b_2)]/[c_2/(c_2+d_2)]$
Study 3	a_3	b_3	c_3	d_3	RR = $[a_3/(a_3+b_3)]/[c_3/(c_3+d_3)]$
Part 2. Usual presentation of a test with 3 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 3 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+$_i$	$(\sum_{i=1}^{n} D+_i) - D+_i$	D-$_i$	$(\sum_{i=1}^{n} D-_i) - D-_i$	$LR_i = \frac{D+_i \big/ \sum_{i=1}^{n} D+_i}{D-_i \big/ \sum_{i=1}^{n} D-_i}$

* Sum of number of patients in moderate and high-risk groups with recurrence, i.e. 103 + 38 = 141 for recurrence group

** Sum of number of patients in low and high-risk groups with recurrence, i.e. 69 + 38 = 107 for recurrence group

+ Sum of number of patients in low and moderate-risk groups with recurrence, i.e. 69 + 103 = 172 for recurrence group

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.

AuthorYear	Year	RiskGroup	RecurInRiskGroup	RecurNotInRiskGroup	NoRecurInRiskGroup	NoRecurNotInRiskGroup	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

Figure 1 legend

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.

References

1. Centor RM, Witherspoon JM, Dalton HP, Brody CE, Link K: **The diagnosis of strep throat in adults in the emergency room**. *Medical decision making : an international journal of the Society for Medical Decision Making* 1981, **1**(3):239-246.
2. Wells PS, Anderson DR, Rodger M, Forgie M, Kearon C, Dreyer J, Kovacs G, Mitchell M, Lewandowski B, Kovacs MJ: **Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis**. *The New England journal of medicine* 2003, **349**(13):1227-1235.
3. Harrington E, Clyne B, Wesseling N, Sandhu H, Armstrong L, Bennett H, Fahey T: **Diagnosing malignant melanoma in ambulatory care: a systematic review of clinical prediction rules**. *BMJ open* 2017, **7**(3):e014096.
4. Ebell MH, Jang W, Shen Y, Geocadin RG, Get With the Guidelines-Resuscitation I: **Development and validation of the Good Outcome Following Attempted Resuscitation (GO-FAR) score to predict neurologically intact survival after in-hospital cardiopulmonary resuscitation**. *JAMA internal medicine* 2013, **173**(20):1872-1878.
5. Pauker SG, Kassirer JP: **The threshold approach to clinical decision making**. *The New England journal of medicine* 1980, **302**(20):1109-1117.
6. Ebell M: **AHRQ White Paper: Use of clinical decision rules for point-of-care decision support**. *Medical decision making : an international journal of the Society for Medical Decision Making* 2010, **30**(6):712-721.
7. van der Helm-van Mil AH, le Cessie S, van Dongen H, Breedveld FC, Toes RE, Huizinga TW: **A prediction rule for disease outcome in patients with recent-onset undifferentiated arthritis: how to guide individual treatment decisions**. *Arthritis and rheumatism* 2007, **56**(2):433-440.
8. Ebell MH, Culp M, Lastinger K, Dasigi T: **A systematic review of the bimanual examination as a test for ovarian cancer**. *American journal of preventive medicine* 2015, **48**(3):350-356.
9. Ebell MH, Culp MB, Radke TJ: **A Systematic Review of Symptoms for the Diagnosis of Ovarian Cancer**. *American journal of preventive medicine* 2016, **50**(3):384-394.
10. Meurs P, Galvin R, Fanning DM, Fahey T: **Prognostic value of the CAPRA clinical prediction rule: a systematic review and meta-analysis**. *BJU international* 2013, **111**(3):427-436.
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**:267-280.
12. Kester AD, Buntinx F: **Meta-analysis of ROC curves**. *Medical decision making : an international journal of the Society for Medical Decision Making* 2000, **20**(4):430-439.
13. van Doorn S, Debray TPA, Kaasenbrood F, Hoes AW, Rutten FH, Moons KGM, Geersing GJ: **Predictive performance of the CHA2DS2-VASc rule in atrial fibrillation: a systematic review and meta-analysis**. *Journal of thrombosis and haemostasis : JTH* 2017, **15**(6):1065-1077.
14. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, Riley RD, Moons KG: **A guide to systematic review and meta-analysis of prediction model performance**. *Bmj* 2017, **356**:i6460.

15. Ebell MH WM, Fahey T, Kearney M, Marchello C: **Meta-analysis of Calibration, Discrimination, and Stratum-Specific Likelihood Ratios for the CRB-65 Score**. *Annals of family medicine* 2019((in press)).

16. Kontopantelis EaR, D.: **metaan: Random-effects meta-analysis**. *The Stata Journal* 2010, **10**(3):395-407.

17. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS: **Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration**. *Annals of internal medicine* 2015, **162**(1):W1-73.

18. Lim WS, van der Eerden MM, Laing R, Boersma WG, Karalus N, Town GI, Lewis SA, Macfarlane JT: **Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study**. *Thorax* 2003, **58**(5):377-382.

19. Deeks JJ, Altman DG: **Diagnostic tests 4: likelihood ratios**. *Bmj* 2004, **329**(7458):168-169.

20. Cooperberg MR, Freedland SJ, Pasta DJ, Elkin EP, Presti JC, Jr., Amling CL, Terris MK, Aronson WJ, Kane CJ, Carroll PR: **Multiinstitutional validation of the UCSF cancer of the prostate risk assessment for prediction of recurrence after radical prostatectomy**. *Cancer* 2006, **107**(10):2384-2391.

21. Brajtbord JS, Leapman MS, Cooperberg MR: **The CAPRA Score at 10 Years: Contemporary Perspectives and Analysis of Supporting Studies**. *European urology* 2017, **71**(5):705-709.

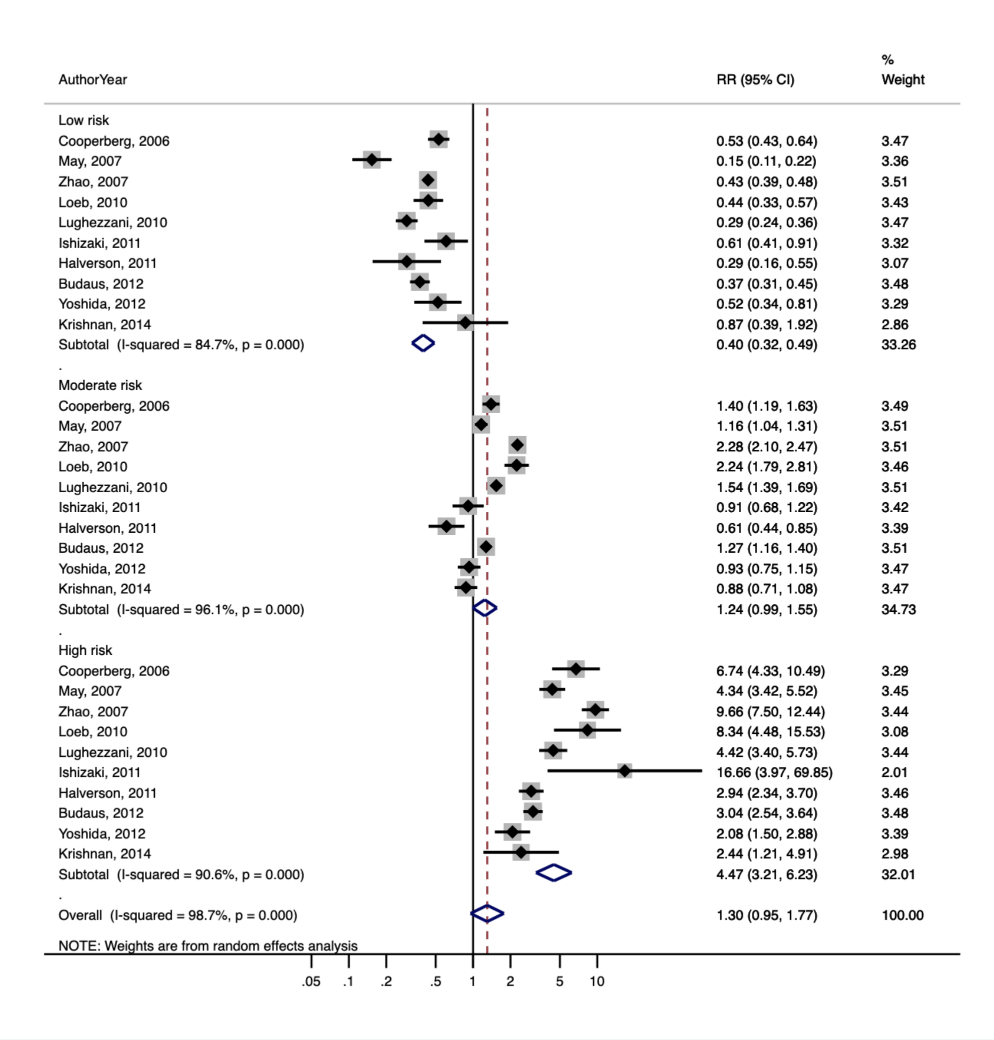
22. Kohn MA, Klok FA, van Es N: **D-dimer Interval Likelihood Ratios for Pulmonary Embolism**. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine* 2017, **24**(7):832-837.

23. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C: **Laboratory diagnosis of iron-deficiency anemia: an overview**. *Journal of general internal medicine* 1992, **7**(2):145-153.

24. Meads C, Ahmed I, Riley RD: **A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance**. *Breast cancer research and treatment* 2012, **132**(2):365-377.

25. Zwinderman AH, Bossuyt PM. **We should not pool diagnostic likelihood ratios in systematic reviews.** *Stat Med*. 2008;27(5):687–697. doi:10.1002/sim.2992

26. 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 Meth Med Res* 2019; 28(9): 2768-86.



AuthorYear	RR (95% CI)	% Weight
Low risk		
Cooperberg, 2006	0.53 (0.43, 0.64)	3.47
May, 2007	0.15 (0.11, 0.22)	3.36
Zhao, 2007	0.43 (0.39, 0.48)	3.51
Loeb, 2010	0.44 (0.33, 0.57)	3.43
Lughezzani, 2010	0.29 (0.24, 0.36)	3.47
Ishizaki, 2011	0.61 (0.41, 0.91)	3.32
Halverson, 2011	0.29 (0.16, 0.55)	3.07
Budaus, 2012	0.37 (0.31, 0.45)	3.48
Yoshida, 2012	0.52 (0.34, 0.81)	3.29
Krishnan, 2014	0.87 (0.39, 1.92)	2.86
Subtotal (I-squared = 84.7%, p = 0.000)	0.40 (0.32, 0.49)	33.26
.		
Moderate risk		
Cooperberg, 2006	1.40 (1.19, 1.63)	3.49
May, 2007	1.16 (1.04, 1.31)	3.51
Zhao, 2007	2.28 (2.10, 2.47)	3.51
Loeb, 2010	2.24 (1.79, 2.81)	3.46
Lughezzani, 2010	1.54 (1.39, 1.69)	3.51
Ishizaki, 2011	0.91 (0.68, 1.22)	3.42
Halverson, 2011	0.61 (0.44, 0.85)	3.39
Budaus, 2012	1.27 (1.16, 1.40)	3.51
Yoshida, 2012	0.93 (0.75, 1.15)	3.47
Krishnan, 2014	0.88 (0.71, 1.08)	3.47
Subtotal (I-squared = 96.1%, p = 0.000)	1.24 (0.99, 1.55)	34.73
.		
High risk		
Cooperberg, 2006	6.74 (4.33, 10.49)	3.29
May, 2007	4.34 (3.42, 5.52)	3.45
Zhao, 2007	9.66 (7.50, 12.44)	3.44
Loeb, 2010	8.34 (4.48, 15.53)	3.08
Lughezzani, 2010	4.42 (3.40, 5.73)	3.44
Ishizaki, 2011	16.66 (3.97, 69.85)	2.01
Halverson, 2011	2.94 (2.34, 3.70)	3.46
Budaus, 2012	3.04 (2.54, 3.64)	3.48
Yoshida, 2012	2.08 (1.50, 2.88)	3.39
Krishnan, 2014	2.44 (1.21, 4.91)	2.98
Subtotal (I-squared = 90.6%, p = 0.000)	4.47 (3.21, 6.23)	32.01
.		
Overall (I-squared = 98.7%, p = 0.000)	1.30 (0.95, 1.77)	100.00

NOTE: Weights are from random effects analysis

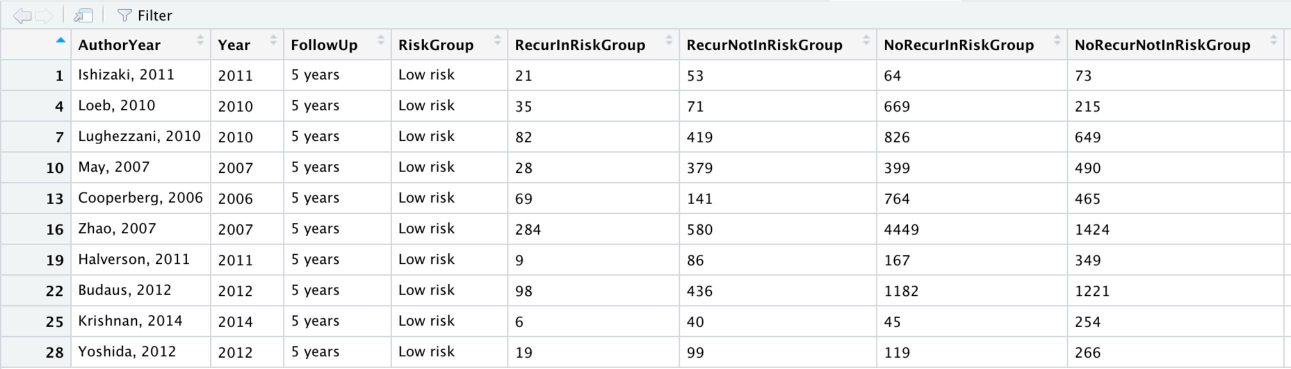
Figure 1. This forest plot shows summary estimates of the stratum specific likelihood ratio for patients classified as low, moderate and high risk for 5-year biochemical recurrence by the CAPRA score.

110x114mm (300 x 300 DPI)

Appendix. Calculating stratum specific likelihood ratios using R

See supplemental files on Zenodo.org to download the R code (https://doi.org/10.5281/zenodo.3936001)

For analysis in R, we loaded the meta library, created a data frame, and then subsets for each risk group. The data set looks like this in R Studio:

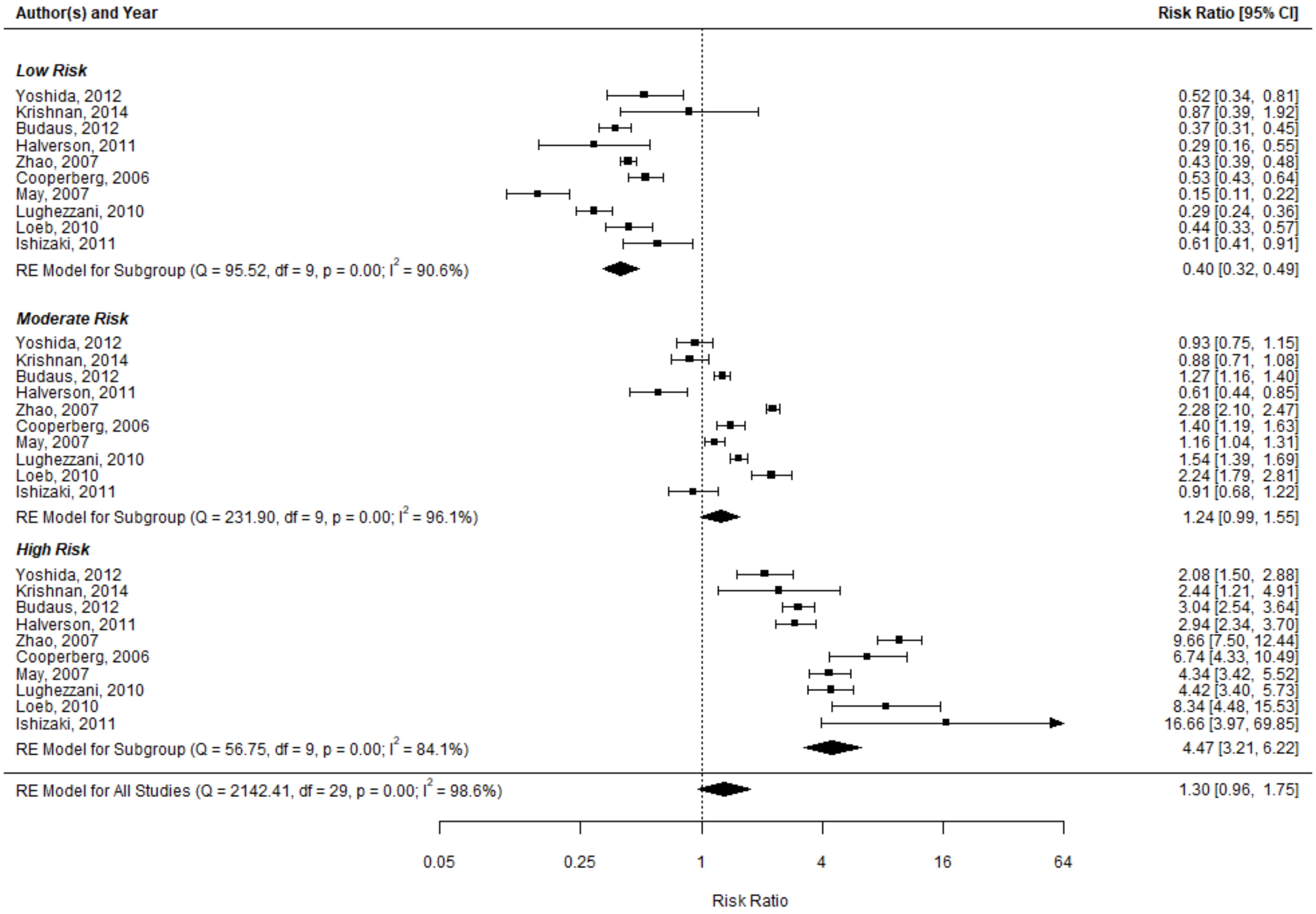


	AuthorYear	Year	FollowUp	RiskGroup	RecurInRiskGroup	RecurNotInRiskGroup	NoRecurInRiskGroup	NoRecurNotInRiskGroup
1	Ishizaki, 2011	2011	5 years	Low risk	21	53	64	73
4	Loeb, 2010	2010	5 years	Low risk	35	71	669	215
7	Lughezzani, 2010	2010	5 years	Low risk	82	419	826	649
10	May, 2007	2007	5 years	Low risk	28	379	399	490
13	Cooperberg, 2006	2006	5 years	Low risk	69	141	764	465
16	Zhao, 2007	2007	5 years	Low risk	284	580	4449	1424
19	Halverson, 2011	2011	5 years	Low risk	9	86	167	349
22	Budaus, 2012	2012	5 years	Low risk	98	436	1182	1221
25	Krishnan, 2014	2014	5 years	Low risk	6	40	45	254
28	Yoshida, 2012	2012	5 years	Low risk	19	99	119	266

```
#import the data
library(readxl)
CAPRA_620 <- read_excel("CAPRA_data_export.xlsx")

# Keep only complete cases since it is importing some empty rows
CAPRA_620<-CAPRA_620[complete.cases(CAPRA_620),]

# Need to make RiskGroup a factor
CAPRA_620$RiskGroup<-as.factor(CAPRA_620$RiskGroup)
# change the order of the factor levels
CAPRA_620$RiskGroup<-factor(CAPRA_620$RiskGroup, levels =c("Low risk","Moderate risk","High risk"))
# Sort the data data by risk group then author name.
CAPRA_620<-CAPRA_620[ with(CAPRA_620, order(-xtfrm(RiskGroup))),]

# I am making the 2x2 table like this.
library(metafor)

#Calculate values using a random effects MH model
res1<-rma.uni(measure="RR", ai= RecurInRiskGroup, bi = RecurNotInRiskGroup,
          ci = NoRecurInRiskGroup, di = NoRecurNotInRiskGroup,
          data = CAPRA_620, method = "DL")


forest(res1, xlim=c(-8, 7), ylim = c(-1,47), at=log(c(0.05, 0.25, 1, 4, 16, 64)), atransf=exp, cex=0.75,
rows=c(3:12,17:26,32:41),
      xlab="Risk Ratio", slab=paste(CAPRA_620$AuthorYear), mlab="", psize=1, header="Author(s) and Year")

### add text with Q-value, dfs, p-value, and I^2 statistic
text(-8, -1, pos=4, cex=0.75, bquote(paste("RE Model for All Studies (Q = ",
                                .(formatC(res1$QE, digits=2, format="f")), ", df = ", .(res1$k - res1$p),
                                ", p = ", .(formatC(res1$QEp, digits=2, format="f")), "; ", I^2, " = ",
                                .(formatC(res1$I2, digits=1, format="f")), "%)")))

### font and save original settings in object 'op'
op <- par(cex=0.75, font=4)

### add text for the subgroups
```

```
text(-8, c(13.5,27.5,42.5), pos=4, c("High Risk",
                              "Moderate Risk",
                              "Low Risk"))

### switch to bold font
par(font=2)

### set par back to the original settings
par(op)

### fit random-effects model in the three subgroups
res.h <- rma.uni(measure="RR", ai= RecurInRiskGroup, bi = RecurNotInRiskGroup,
             ci = NoRecurInRiskGroup, di = NoRecurNotInRiskGroup,
             subset=(RiskGroup=="High risk"), data = CAPRA_620,
             method = "DL")
res.m <- rma.uni(measure="RR", ai= RecurInRiskGroup, bi = RecurNotInRiskGroup,
             ci = NoRecurInRiskGroup, di = NoRecurNotInRiskGroup,
             subset=(RiskGroup=="Moderate risk"), data = CAPRA_620,
             method = "DL")
res.l <- rma.uni(measure="RR", ai= RecurInRiskGroup, bi = RecurNotInRiskGroup,
             ci = NoRecurInRiskGroup, di = NoRecurNotInRiskGroup,
             subset=(RiskGroup=="Low risk"), data = CAPRA_620,
             method = "DL")

### add summary polygons for the three subgroups
addpoly(res.h, row= 1.5, cex=0.75, atransf=exp, mlab="")
addpoly(res.m, row= 15.5, cex=0.75, atransf=exp, mlab="")
addpoly(res.l, row= 30.5, cex=0.75, atransf=exp, mlab="")

### add text with Q-value, dfs, p-value, and I^2 statistic for subgroups
text(-8, 30.5, pos=4, cex=0.75, bquote(paste("RE Model for Subgroup (Q = ",
                                     .(formatC(res.h$QE, digits=2, format="f")), ", df = ", .(res.h$k - res.h$p),
                                     ", p = ", .(formatC(res.h$QEp, digits=2, format="f")), "; ", I^2, " = ",
                                     .(formatC(res.h$I2, digits=1, format="f")), "%)")))
text(-8, 15.5, pos=4, cex=0.75, bquote(paste("RE Model for Subgroup (Q = ",
                                     .(formatC(res.m$QE, digits=2, format="f")), ", df = ", .(res.m$k - res.m$p),
                                     ", p = ", .(formatC(res.m$QEp, digits=2, format="f")), "; ", I^2, " = ",
                                     .(formatC(res.m$I2, digits=1, format="f")), "%)")))
text(-8, 1.5, pos=4, cex=0.75, bquote(paste("RE Model for Subgroup (Q = ",
                                    .(formatC(res.l$QE, digits=2, format="f")), ", df = ", .(res.l$k - res.l$p),
                                    ", p = ", .(formatC(res.l$QEp, digits=2, format="f")), "; ", I^2, " = ",
                                    .(formatC(res.l$I2, digits=1, format="f")), "%)")))
```

The output is shown below:


Author(s) and Year
Risk Ratio [95% CI]
Low Risk
Yoshida, 2012 0.52 [0.34, 0.81]
Krishnan, 2014 0.87 [0.39, 1.92]
Budaus, 2012 0.37 [0.31, 0.45]
Halverson, 2011 0.29 [0.16, 0.55]
Zhao, 2007 0.43 [0.39, 0.48]
Cooperberg, 2006 0.53 [0.43, 0.64]
May, 2007 0.15 [0.11, 0.22]
Lughezzani, 2010 0.29 [0.24, 0.36]
Loeb, 2010 0.44 [0.33, 0.57]
Ishizaki, 2011 0.61 [0.41, 0.91]
RE Model for Subgroup (Q = 95.52, df = 9, p = 0.00; I^2 = 90.6%) 0.40 [0.32, 0.49]
Moderate Risk
Yoshida, 2012 0.93 [0.75, 1.15]
Krishnan, 2014 0.88 [0.71, 1.08]
Budaus, 2012 1.27 [1.16, 1.40]
Halverson, 2011 0.61 [0.44, 0.85]
Zhao, 2007 2.28 [2.10, 2.47]
Cooperberg, 2006 1.40 [1.19, 1.63]
May, 2007 1.16 [1.04, 1.31]
Lughezzani, 2010 1.54 [1.39, 1.69]
Loeb, 2010 2.24 [1.79, 2.81]
Ishizaki, 2011 0.91 [0.68, 1.22]
RE Model for Subgroup (Q = 231.90, df = 9, p = 0.00; I^2 = 96.1%) 1.24 [0.99, 1.55]
High Risk
Yoshida, 2012 2.08 [1.50, 2.88]
Krishnan, 2014 2.44 [1.21, 4.91]
Budaus, 2012 3.04 [2.54, 3.64]
Halverson, 2011 2.94 [2.34, 3.70]
Zhao, 2007 9.66 [7.50, 12.44]
Cooperberg, 2006 6.74 [4.33, 10.49]
May, 2007 4.34 [3.42, 5.52]
Lughezzani, 2010 4.42 [3.40, 5.73]
Loeb, 2010 8.34 [4.48, 15.53]
Ishizaki, 2011 16.66 [3.97, 69.85]
RE Model for Subgroup (Q = 56.75, df = 9, p = 0.00; I^2 = 84.1%) 4.47 [3.21, 6.22]
RE Model for All Studies (Q = 2142.41, df = 29, p = 0.00; I^2 = 98.6%) 1.30 [0.96, 1.75]
0.05
0.25
1
4
16
64
Risk Ratio

BMJ Open

A Novel Approach to Meta-Analysis of Tests and Clinical Prediction Rules with 3 or More Risk Categories

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-036262.R2
Article Type:	Original research
Date Submitted by the Author:	09-Nov-2020
Complete List of Authors:	Ebell, Mark; University of Georgia, Epidemiology and Biostatistics Walsh, Mary; Royal College of Surgeons in Ireland, HRB Centre for Primary Care Research, Department of General Practice; Royal College of Surgeons in Ireland, School of Physiotherapy Boland, Fiona; Royal College of Surgeons in Ireland, HRB Centre For Primary Care Research, Division of Population Health Sciences (PHS) McKay, Brian; University of Georgia, Epidemiology and Biostatistics Fahey, Tom; Royal College of Surgeons in Ireland, Department of General Practice
<b>Primary Subject Heading</b>:	Evidence based practice
Secondary Subject Heading:	Diagnostics, General practice / Family practice, Health services research, Oncology
Keywords:	PRIMARY CARE, STATISTICS & RESEARCH METHODS, Urological tumours < ONCOLOGY

A Novel Approach to Meta-Analysis of Tests and clinical prediction rules with 3 or more risk categories

Mark H. Ebell (1)

Mary E. Walsh (2)

Fiona Boland (2)

Brian McKay (1)

Tom Fahey (2)

1. Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia, Athens, GA, USA. 2. HRB Centre for Primary Care Research, Royal College of Surgeons in Ireland, 123 St. Stephens Green, Dublin 2, Ireland

Corresponding Author:

Mark H. Ebell MD, MS
125 B.S. Miller Hall
UGA Health Sciences Campus
Athens, GA 30602
706-247-4953
ebell@uga.edu

Word count: 3778 words (not including abstract or references), 3 tables and 1 figure

Keywords: meta-analysis, likelihood ratio, diagnostic accuracy, systematic review, multichotomous

Abstract

Objective

Multichotomous tests have 3 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 dichotomize 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 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 0.32-0.49), 1.24 (95% CI 0.99-1.55), and 4.47 (95% CI 3.21-6.23) 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

- 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 3 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 ABCD rule for the evaluation of skin lesions,[3] and the 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 dichotomizing, 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) one year later among patients presenting with undifferentiated joint pain to a general practitioner.[7] Simply dichotomizing the risk score into low and high risk groups based on a single cutoff that maximizes 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 2 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 cutoff. Instead, the preferred measure of diagnostic accuracy for multichotomous tests and CPRs is the stratum specific likelihood ratio, i.e. 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 over-prediction of the diagnosis, and a RR < 1.0 under-prediction;[10, 11]
3) performing meta-analysis of receiver operating characteristic curves;[12]
4) dichotomizing the test, by combining groups until there are only two dichotomous categories with a single cutoff, 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, dichotomizing CPRs that have 3 or more risk groups into 2 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 dichotomize 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 CAPRA score for prostate cancer prognosis.

Methods

Calculating Summary Estimates of Stratum Specific Likelihood Ratios (SSLR)

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 likelihood ratios are commonly reported. For a multichotomous test or CPR with more 3 or more risk categories, each risk category has its own likelihood ratio, called the “stratum specific likelihood ratio” (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 SSRL, 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 randomized 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 cutoffs 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 likelihood ratio. 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 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 CPRs the outcome being predicted is often the likelihood of disease, hence use of D):

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

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

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

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 risk ratio (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 risk ratios for the calculation of SSLRs.

Part 1 shows how data are formatted for a meta-analysis of 3 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 3 or more risk groups, and how the stratum specific likelihood ratios for a single study are calculated. Part 3 reformats the same data to mimic the risk ratios of a treatment trial, illustrating how the risk ratios are identical to the likelihood ratios calculated in Part 2. Finally, Part 4 illustrates the general case for formatting the results of a study describing a CPR with 3 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 cutoff, 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 TX) we used the metan procedure (version 9) to perform a random effects meta-analysis of risk ratios 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 Appendix. SAS has no intrinsic features for meta-analysis. Prof. Stephen Senn and colleagues produced a suite of detailed macros, which can be downloaded from: http://www.senns.demon.co.uk/SAS%20Macros/SASMacros.html.

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 (AUC) and corresponding measures of uncertainty (95% confidence intervals or standard errors) can be extracted from individual studies, where reported. In order to conduct meta-analysis, AUC values and reported 95% confidence intervals 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 REML estimation, which can be completed for example using the metaan procedure in Stata 16 (Stata Corp, College Station

TX).[14,20] The pooled logit AUC and 95% confidence intervals 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 (observed: expected [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 odds ratio 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 (Stata Corp, College Station TX).[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 area under the curve 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 likelihood ratios 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 risk ratios 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 0.32-0.49) for the low risk group, 1.24 (95% CI 0.99-1.55) for the moderate risk group, and 4.47 (95% CI 3.21-6.23) 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 likelihood ratios 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 labeled “Risk Ratio”, although this could easily be modified using a graphics program (development of a native R package is underway).

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 2 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 stratum specific likelihood ratios.[11,14] It does not require dichotomizing 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 3 or more risk categories for an outcome, our approach to the calculation of summary estimates of SSLR could also be applied to any multichotomous diagnostic test such as serum ferritin or d-dimer.[23,24]

Zwinderman and Bossuyt argue that meta-analysis of diagnostic likelihood ratios is not appropriate, since the positive and negative likelihood ratios are highly correlated for a dichotomous test, because they are calculated from sensitivity and specificity which are also highly correlated.[25] Therefore, they suggest that bivariate meta-analysis of sensitivity and specificity should be performed instead of meta-analysis of likelihood ratios, with subsequent calculation of positive and negative likelihood ratios if desired. However, this is not relevant for stratum specific likelihood ratios 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 stratum specific likelihood ratios (without dichotomizing 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 0.97-1.20).[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 likelihood ratios is that while in theory likelihood ratios 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 likelihood ratios between studies.[27] By using a random effects model in our meta-analysis of stratum specific likelihood ratios, 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 confidence interval 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 stratum specific likelihood ratios for any test with 3 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.

Contributorship Statement

The project was conceptualized and led by Mark Ebell. Brian McKay wrote and tested the R code. Tom Fahey provided input on the conceptualization, assisted with writing, and reviewed the final manuscript. Mary Walsh and Fiona Boland collaborated on the meta-analysis of ROC 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.

Acknowledgement

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 CPRs.

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.

Data sharing agreement

There was no original data collection for this study. R code and anexcel 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

Competing interests statement

The authors have no competing interests either financial or intellectual to disclose.

Table 1. Calculation of stratum specific likelihood ratios for a single study of the 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	

Table 2. Developing a method for formatting data from tests or clinical decision rules with 3 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 = $[a_1/(a_1+b_1)]/[c_1/(c_1+d_1)]$
Study 2	a_2	b_2	c_2	d_2	RR = $[a_2/(a_2+b_2)]/[c_2/(c_2+d_2)]$
Study 3	a_3	b_3	c_3	d_3	RR = $[a_3/(a_3+b_3)]/[c_3/(c_3+d_3)]$
Part 2. Usual presentation of a test with 3 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 3 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+$_i$	$(\sum_{i=1}^{n} D+_i) - D+_i$	D-$_i$	$(\sum_{i=1}^{n} D-_i) - D-_i$	$LR_i = \frac{D+_i / \sum_{i=1}^{n} D+_i}{D-_i / \sum_{i=1}^{n} D-_i}$

* Sum of number of patients in moderate and high-risk groups with recurrence, i.e. 103 + 38 = 141 for recurrence group

** Sum of number of patients in low and high-risk groups with recurrence, i.e. 69 + 38 = 107 for recurrence group

+ Sum of number of patients in low and moderate-risk groups with recurrence, i.e. 69 + 103 = 172 for recurrence group

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.

AuthorYear	Year	RiskGroup	RecurInRiskGroup	RecurNotInRiskGroup	NoRecurInRiskGroup	NoRecurNotInRiskGroup	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

Figure 1 legend

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.

References

1. Centor RM, Witherspoon JM, Dalton HP, Brody CE, Link K: The diagnosis of strep throat in adults in the emergency room. Med Decis Making 1981, 1(3):239-246.
2. Wells PS, Anderson DR, Rodger M, Forgie M, Kearon C, Dreyer J, Kovacs G, Mitchell M, Lewandowski B, Kovacs MJ: Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis. N Engl J Med 2003, 349(13):1227-1235.
3. Harrington E, Clyne B, Wesseling N, Sandhu H, Armstrong L, Bennett H, Fahey T: Diagnosing malignant melanoma in ambulatory care: a systematic review of clinical prediction rules. BMJ Open 2017, 7(3):e014096.
4. Ebell MH, Jang W, Shen Y, Geocadin RG, Get With the Guidelines-Resuscitation I: Development and validation of the Good Outcome Following Attempted Resuscitation (GO-FAR) score to predict neurologically intact survival after in-hospital cardiopulmonary resuscitation. JAMA Int Med 2013, 173(20):1872-1878.
5. Pauker SG, Kassirer JP: The threshold approach to clinical decision making. N Engl J Med 1980, 302(20):1109-1117.
6. Ebell M: AHRQ White Paper: Use of clinical decision rules for point-of-care decision support. Med Decis Making 2010, 30(6):712-721.
7. van der Helm-van Mil AH, le Cessie S, van Dongen H, Breedveld FC, Toes RE, Huizinga TW: A prediction rule for disease outcome in patients with recent-onset undifferentiated arthritis: how to guide individual treatment decisions. Arthritis and Rheumatism 2007, 56(2):433-440.
8. Ebell MH, Culp M, Lastinger K, Dasigi T: A systematic review of the bimanual examination as a test for ovarian cancer. Am J Prev Med 2015, 48(3):350-356.
9. Ebell MH, Culp MB, Radke TJ: A Systematic Review of Symptoms for the Diagnosis of Ovarian Cancer. American J Prev Med 2016, 50(3):384-394.
10. Meurs P, Galvin R, Fanning DM, Fahey T: Prognostic value of the CAPRA clinical prediction rule: a systematic review and meta-analysis. BJU international 2013, 111(3):427-436.
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, 7:267-280.
12. Kester AD, Buntinx F: Meta-analysis of ROC curves. Med Decis Making 2000, 20(4):430-439.
13. van Doorn S, Debray TPA, Kaasenbrood F, Hoes AW, Rutten FH, Moons KGM, Geersing GJ: Predictive performance of the CHA2DS2-VASc rule in atrial fibrillation: a systematic review and meta-analysis. J Thrombosis Haemostasis : JTH 2017, 15(6):1065-1077.
14. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, Riley RD, Moons KG: 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 Meth Med Res 2019; 28(9): 2768-86.
16. Deeks JJ, Altman DG: Diagnostic tests 4: likelihood ratios. BMJ 2004, 329(7458):168-169.
17. Cooperberg MR, Freedland SJ, Pasta DJ, Elkin EP, Presti JC, Jr., Amling CL, Terris MK, Aronson WJ, Kane CJ, Carroll PR: Multiinstitutional validation of the UCSF cancer of the prostate risk assessment for prediction of recurrence after radical prostatectomy. Cancer 2006, 107(10):2384-2391.

18. Brajtbord JS, Leapman MS, Cooperberg MR: The CAPRA Score at 10 Years: Contemporary Perspectives and Analysis of Supporting Studies. Eur Urol 2017, 71(5):705-709.
19. Ebell MH WM, Fahey T, Kearney M, Marchello C: Meta-analysis of Calibration, Discrimination, and Stratum-Specific Likelihood Ratios for the CRB-65 Score. J Gen Intern Med 2019; 34(7): 1304-1313.
20. Kontopantelis EaR, D.: metaan: Random-effects meta-analysis. The Stata Journal 2010, 10(3):395-407.
21. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015, 162(1):W1-73.
22. Lim WS, van der Eerden MM, Laing R, Boersma WG, Karalus N, Town GI, Lewis SA, Macfarlane JT: Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax 2003, 58(5):377-382.
23. Kohn MA, Klok FA, van Es N: D-dimer Interval Likelihood Ratios for Pulmonary Embolism. Acad Emerg Med 2017, 24(7):832-837.
24. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C: Laboratory diagnosis of iron-deficiency anemia: an overview. J Gen Intern Med 1992, 7(2):145-153.
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(2):365-377.
25. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. Stat Med. 2008;27(5):687–697. doi:10.1002/sim.2992
27. Leeflang M, Rutjes A, Reitsma J, et al. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ 2013; 185 (11) E537-E544

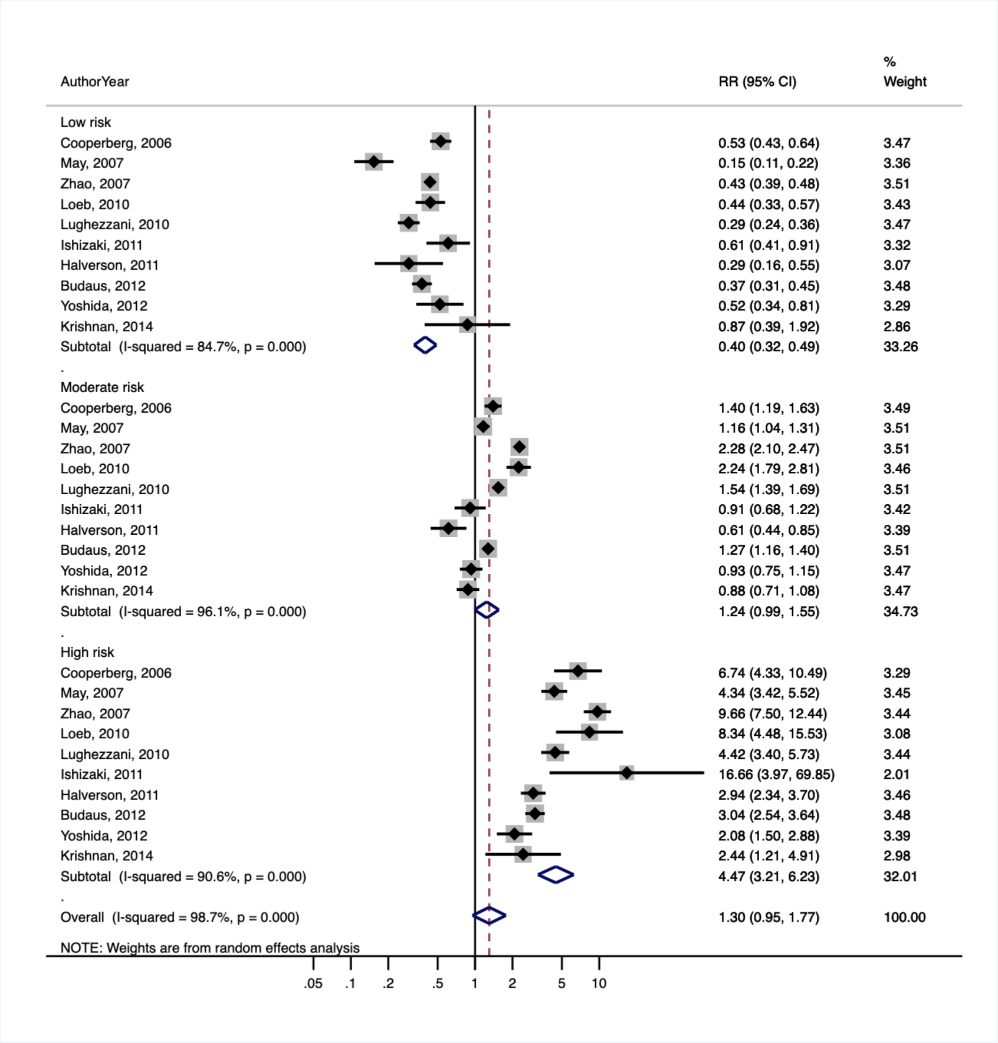


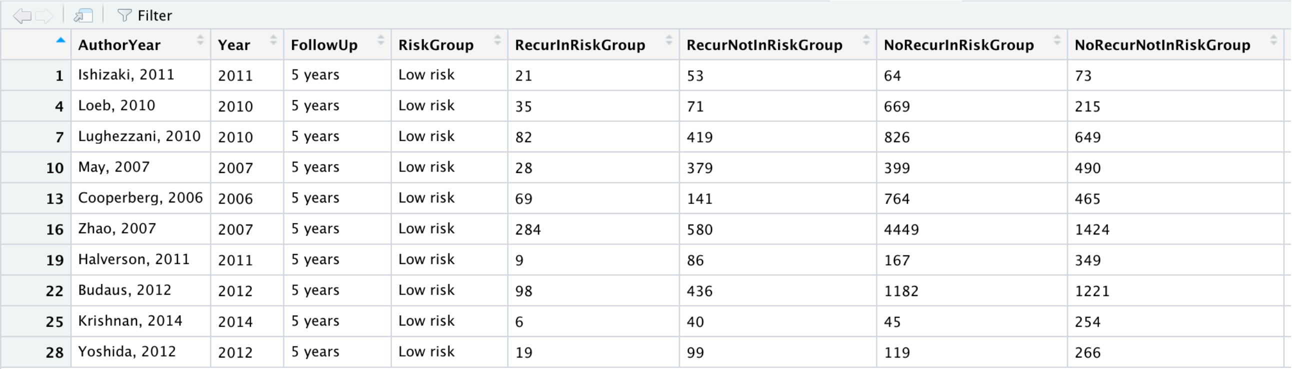
Figure 1. This forest plot shows summary estimates of the stratum specific likelihood ratio for patients classified as low, moderate and high risk for 5-year biochemical recurrence by the CAPRA score.

110x114mm (300 x 300 DPI)

Appendix. Calculating stratum specific likelihood ratios using R

See supplemental files on Zenodo.org to download the R code (https://doi.org/10.5281/zenodo.3936001)

For analysis in R, we loaded the meta library, created a data frame, and then subsets for each risk group. The data set looks like this in R Studio:

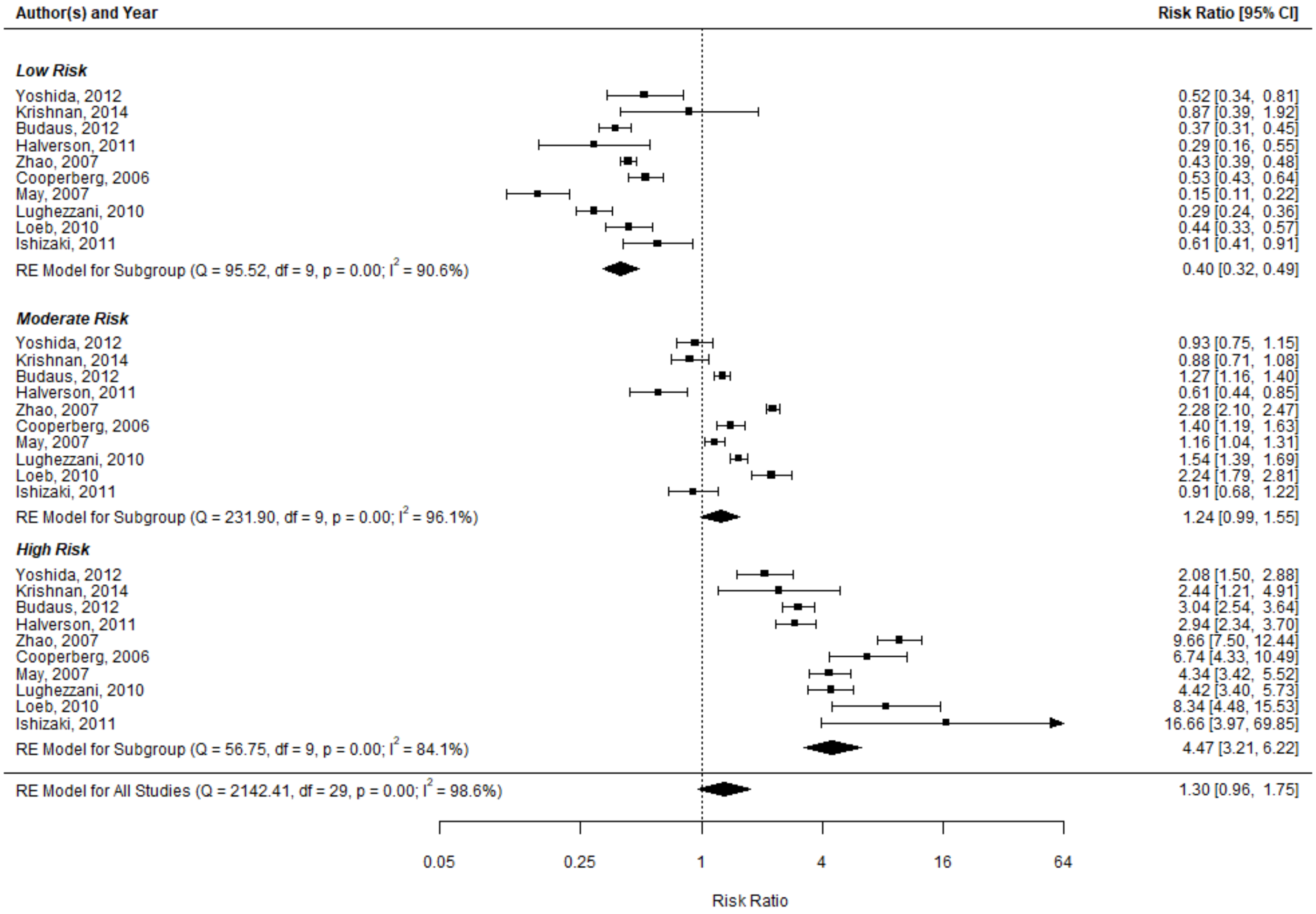


	AuthorYear	Year	FollowUp	RiskGroup	RecurInRiskGroup	RecurNotInRiskGroup	NoRecurInRiskGroup	NoRecurNotInRiskGroup
1	Ishizaki, 2011	2011	5 years	Low risk	21	53	64	73
4	Loeb, 2010	2010	5 years	Low risk	35	71	669	215
7	Lughezzani, 2010	2010	5 years	Low risk	82	419	826	649
10	May, 2007	2007	5 years	Low risk	28	379	399	490
13	Cooperberg, 2006	2006	5 years	Low risk	69	141	764	465
16	Zhao, 2007	2007	5 years	Low risk	284	580	4449	1424
19	Halverson, 2011	2011	5 years	Low risk	9	86	167	349
22	Budaus, 2012	2012	5 years	Low risk	98	436	1182	1221
25	Krishnan, 2014	2014	5 years	Low risk	6	40	45	254
28	Yoshida, 2012	2012	5 years	Low risk	19	99	119	266

```
#import the data
library(readxl)
CAPRA_620 <- read_excel("CAPRA_data_export.xlsx")

# Keep only complete cases since it is importing some empty rows
CAPRA_620<-CAPRA_620[complete.cases(CAPRA_620),]

# Need to make RiskGroup a factor
CAPRA_620$RiskGroup<-as.factor(CAPRA_620$RiskGroup)
# change the order of the factor levels
CAPRA_620$RiskGroup<-factor(CAPRA_620$RiskGroup, levels =c("Low risk","Moderate risk","High risk"))
# Sort the data data by risk group then author name.
CAPRA_620<-CAPRA_620[ with(CAPRA_620, order(-xtfrm(RiskGroup))),]

# I am making the 2x2 table like this.
library(metafor)

#Calculate values using a random effects MH model
res1<-rma.uni(measure="RR", ai= RecurInRiskGroup, bi = RecurNotInRiskGroup,
          ci = NoRecurInRiskGroup, di = NoRecurNotInRiskGroup,
          data = CAPRA_620, method = "DL")


forest(res1, xlim=c(-8, 7), ylim = c(-1,47), at=log(c(0.05, 0.25, 1, 4, 16, 64)), atransf=exp, cex=0.75,
rows=c(3:12,17:26,32:41),
      xlab="Risk Ratio", slab=paste(CAPRA_620$AuthorYear), mlab="", psize=1, header="Author(s) and Year")

### add text with Q-value, dfs, p-value, and I^2 statistic
text(-8, -1, pos=4, cex=0.75, bquote(paste("RE Model for All Studies (Q = ",
                                   .(formatC(res1$QE, digits=2, format="f")), ", df = ", .(res1$k - res1$p),
                                   ", p = ", .(formatC(res1$QEp, digits=2, format="f")), "; ", I^2, " = ",
                                   .(formatC(res1$I2, digits=1, format="f")), "%)")))

### font and save original settings in object 'op'
op <- par(cex=0.75, font=4)

### add text for the subgroups
```

```
text(-8, c(13.5,27.5,42.5), pos=4, c("High Risk",
                                  "Moderate Risk",
                                  "Low Risk"))

### switch to bold font
par(font=2)

### set par back to the original settings
par(op)

### fit random-effects model in the three subgroups
res.h <- rma.uni(measure="RR", ai= RecurInRiskGroup, bi = RecurNotInRiskGroup,
               ci = NoRecurInRiskGroup, di = NoRecurNotInRiskGroup,
               subset=(RiskGroup=="High risk"), data = CAPRA_620,
               method = "DL")
res.m <- rma.uni(measure="RR", ai= RecurInRiskGroup, bi = RecurNotInRiskGroup,
               ci = NoRecurInRiskGroup, di = NoRecurNotInRiskGroup,
               subset=(RiskGroup=="Moderate risk"), data = CAPRA_620,
               method = "DL")
res.l <- rma.uni(measure="RR", ai= RecurInRiskGroup, bi = RecurNotInRiskGroup,
               ci = NoRecurInRiskGroup, di = NoRecurNotInRiskGroup,
               subset=(RiskGroup=="Low risk"), data = CAPRA_620,
               method = "DL")

### add summary polygons for the three subgroups
addpoly(res.h, row= 1.5, cex=0.75, atransf=exp, mlab="")
addpoly(res.m, row= 15.5, cex=0.75, atransf=exp, mlab="")
addpoly(res.l, row= 30.5, cex=0.75, atransf=exp, mlab="")

### add text with Q-value, dfs, p-value, and I^2 statistic for subgroups
text(-8, 30.5, pos=4, cex=0.75, bquote(paste("RE Model for Subgroup (Q = ",
                                         .(formatC(res.h$QE, digits=2, format="f")), ", df = ", .(res.h$k - res.h$p),
                                         ", p = ", .(formatC(res.h$QEp, digits=2, format="f")), "; ", I^2, " = ",
                                         .(formatC(res.h$I2, digits=1, format="f")), "%)")))
text(-8, 15.5, pos=4, cex=0.75, bquote(paste("RE Model for Subgroup (Q = ",
                                         .(formatC(res.m$QE, digits=2, format="f")), ", df = ", .(res.m$k - res.m$p),
                                         ", p = ", .(formatC(res.m$QEp, digits=2, format="f")), "; ", I^2, " = ",
                                         .(formatC(res.m$I2, digits=1, format="f")), "%)")))
text(-8, 1.5, pos=4, cex=0.75, bquote(paste("RE Model for Subgroup (Q = ",
                                        .(formatC(res.l$QE, digits=2, format="f")), ", df = ", .(res.l$k - res.l$p),
                                        ", p = ", .(formatC(res.l$QEp, digits=2, format="f")), "; ", I^2, " = ",
                                        .(formatC(res.l$I2, digits=1, format="f")), "%)")))
```

The output is shown below:


Author(s) and Year
Risk Ratio [95% CI]
Low Risk
Yoshida, 2012
0.52 [0.34, 0.81]
Krishnan, 2014
0.87 [0.39, 1.92]
Budaus, 2012
0.37 [0.31, 0.45]
Halverson, 2011
0.29 [0.16, 0.55]
Zhao, 2007
0.43 [0.39, 0.48]
Cooperberg, 2006
0.53 [0.43, 0.64]
May, 2007
0.15 [0.11, 0.22]
Lughezzani, 2010
0.29 [0.24, 0.36]
Loeb, 2010
0.44 [0.33, 0.57]
Ishizaki, 2011
0.61 [0.41, 0.91]
RE Model for Subgroup (Q = 95.52, df = 9, p = 0.00; I² = 90.6%)
0.40 [0.32, 0.49]
Moderate Risk
Yoshida, 2012
0.93 [0.75, 1.15]
Krishnan, 2014
0.88 [0.71, 1.08]
Budaus, 2012
1.27 [1.16, 1.40]
Halverson, 2011
0.61 [0.44, 0.85]
Zhao, 2007
2.28 [2.10, 2.47]
Cooperberg, 2006
1.40 [1.19, 1.63]
May, 2007
1.16 [1.04, 1.31]
Lughezzani, 2010
1.54 [1.39, 1.69]
Loeb, 2010
2.24 [1.79, 2.81]
Ishizaki, 2011
0.91 [0.68, 1.22]
RE Model for Subgroup (Q = 231.90, df = 9, p = 0.00; I² = 96.1%)
1.24 [0.99, 1.55]
High Risk
Yoshida, 2012
2.08 [1.50, 2.88]
Krishnan, 2014
2.44 [1.21, 4.91]
Budaus, 2012
3.04 [2.54, 3.64]
Halverson, 2011
2.94 [2.34, 3.70]
Zhao, 2007
9.66 [7.50, 12.44]
Cooperberg, 2006
6.74 [4.33, 10.49]
May, 2007
4.34 [3.42, 5.52]
Lughezzani, 2010
4.42 [3.40, 5.73]
Loeb, 2010
8.34 [4.48, 15.53]
Ishizaki, 2011
16.66 [3.97, 69.85]
RE Model for Subgroup (Q = 56.75, df = 9, p = 0.00; I² = 84.1%)
4.47 [3.21, 6.22]
RE Model for All Studies (Q = 2142.41, df = 29, p = 0.00; I² = 98.6%)
1.30 [0.96, 1.75]
0.05
0.25
1
4
16
64
Risk Ratio