Automated quality assessment of medical evidence

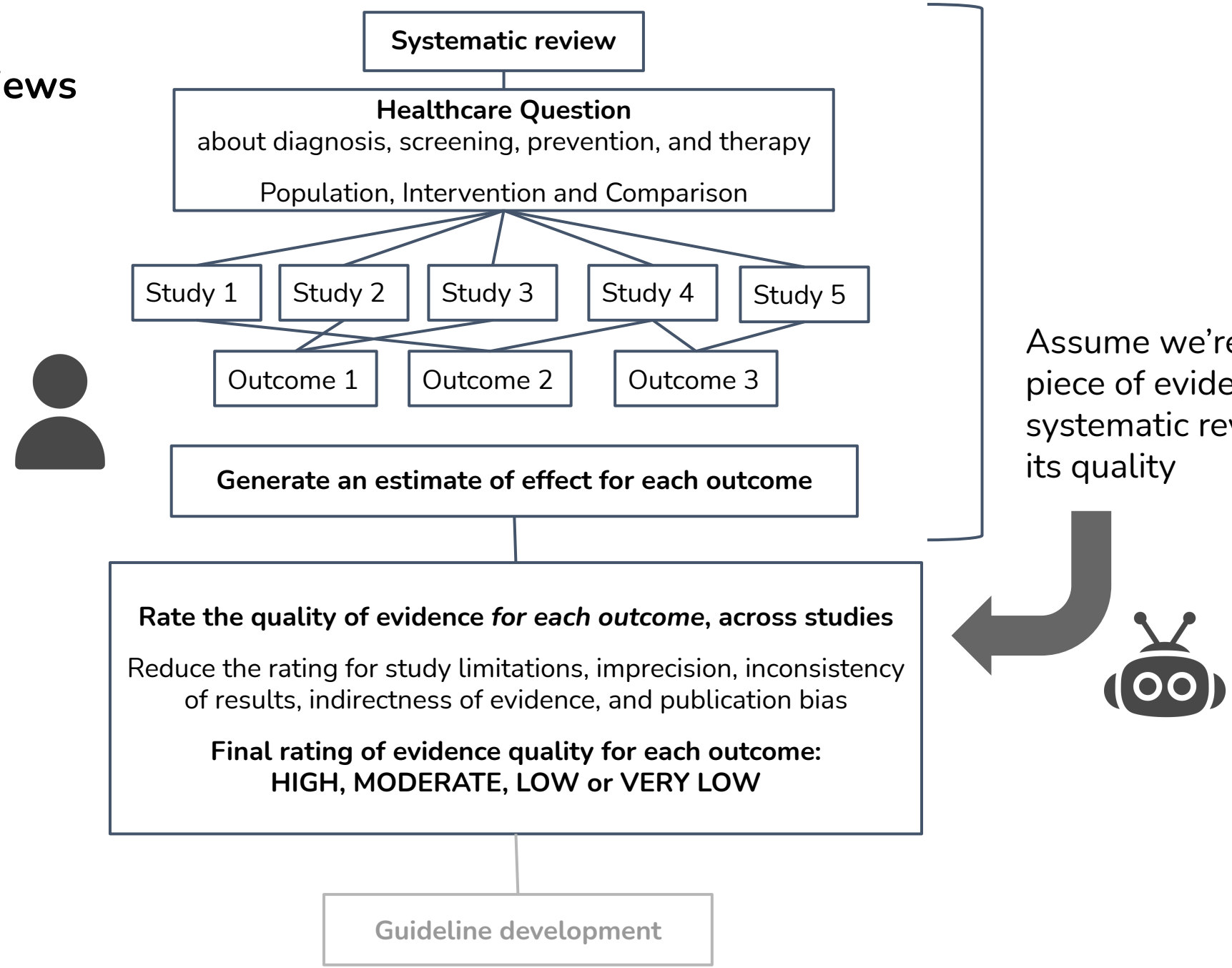
Simon Šuster

with Tim Baldwin, Antonio Jimeno Yepes, Jey Han Lau, David Martinez Iraola, Yulia Otmakhova, Karin Verspoor

Stream 4

22/3/2021

Constructing systematic reviews and quality assessment



Assume we're given a piece of evidence from a systematic review, predict its quality

Modified from Guyatt et al. 2011

Landscape

Predicting strength of recommendation of a body of evidence (Sarker et al., 2015)

- 1,100 abstracts, 3 levels according to SORT (Ebell et al., 2004)
- Publication metadata features and word n-grams: 64% accuracy

Limitations:

- Unclear what the score measures (strongly reflects the publication types)
- Loosely defined SORT criteria and inclusion criteria (doesn't follow PICO)
- Cohen's kappa of around 0.5 for human annotators

Grading individual studies with isolated criteria

- Risk-of-bias assessment in RobotReviewer and TrialStreamer (Marshall et al., 2017 & 2020)

Limitations:

- Does not grade the body of evidence but individual studies

Semi-automated quality assessment (SAQAT; Stewart et al., 2015)

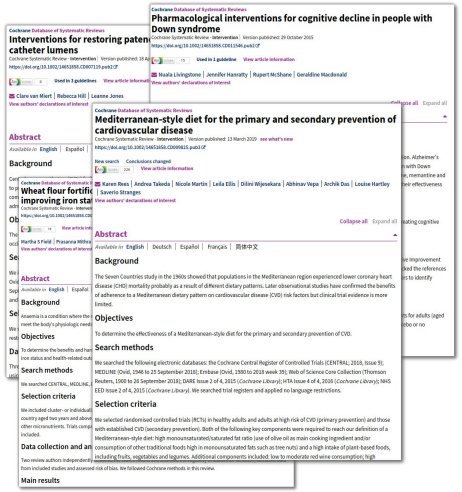
- Human reviewers answer checklist questions
- Final score assigned by a Bayesian network

Limitations:

- Still largely manual

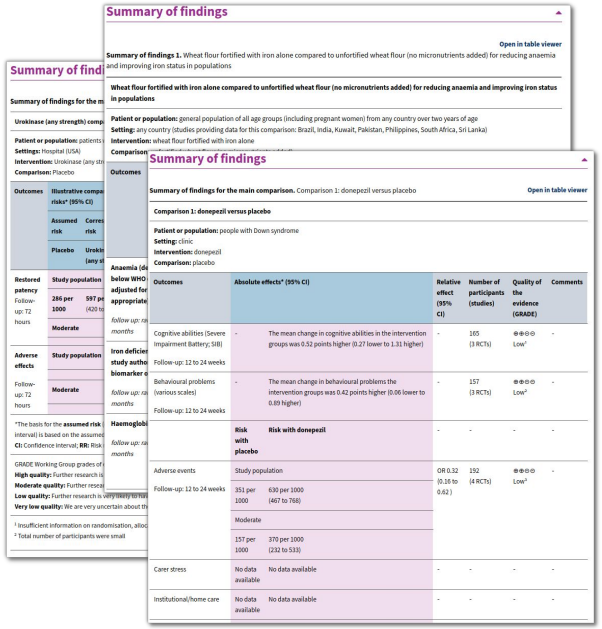
Data creation

Cochrane Library



~7,000 systematic reviews (majority from 2010-)

Extract data related to quality appraisal from summaries of findings



13,500 outcomes rated for quality using GRADE framework (with justifications)

- Validation of extraction procedures against human-verified data (Conway et al., 2017)
- Prepared data for 10-fold CV, with train/dev/test splits
  - instances built from same SR kept in the same split
- Author-assigned quality scores represent our gold standard (labels)

Levosimendan compared with placebo for cardiogenic shock or low cardiac output syndrome

Patient or population: adults with cardiogenic shock or low cardiac output syndrome

Settings: hospital

Intervention: levosimendan

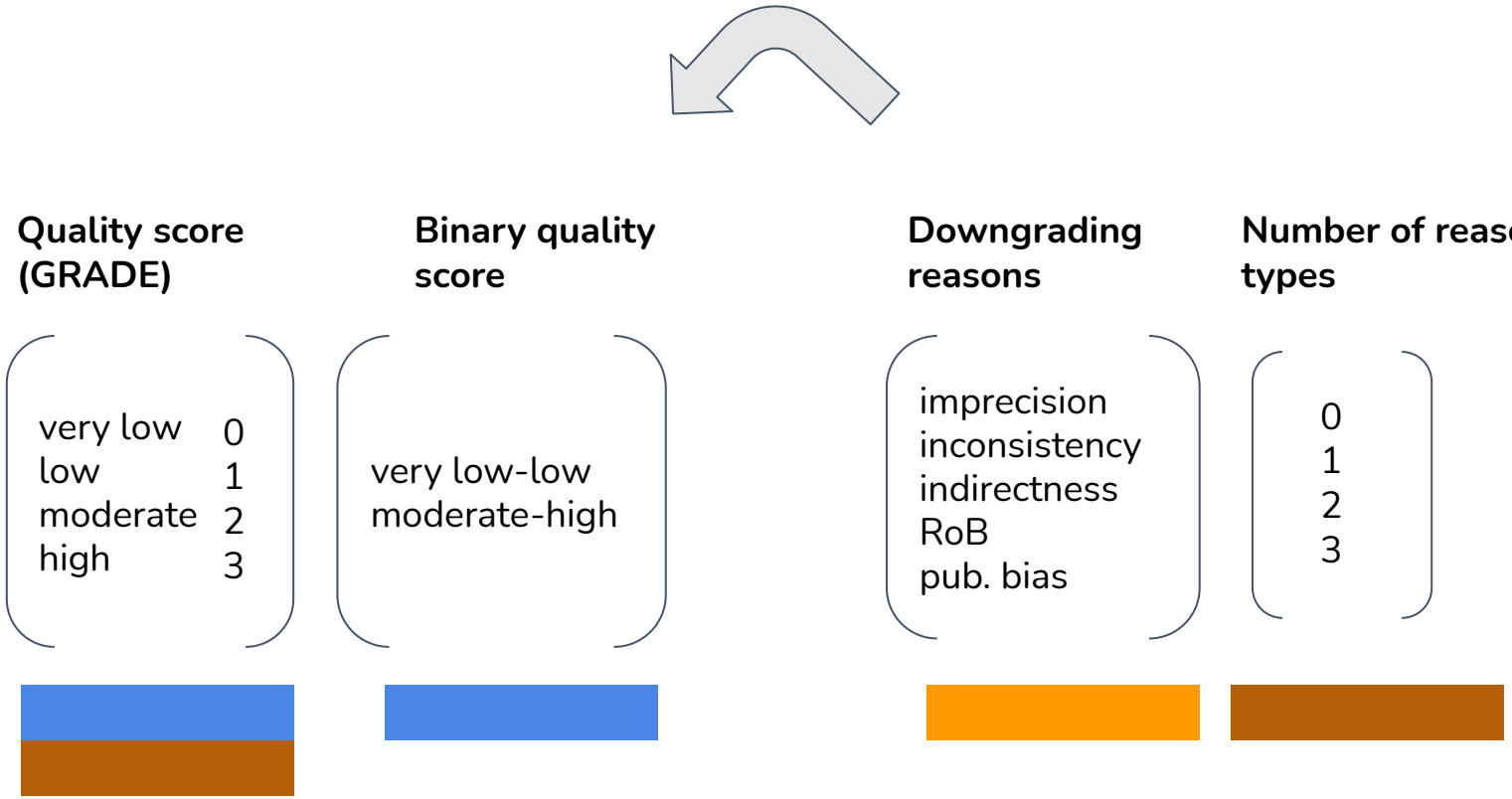
Comparison: placebo

Outcomes	Anticipated absolute effects (95% CI)		Relative effect (95% CI)	No of participants (studies)	Quality	Comments
	Risk with placebo	Risk with levosimendan				
All-cause short-term mortality: range 4 to 6 months	**Moderate**[1]		**RR 0.48** (0.12 to 1.94)	55 (2)	⊕⊕⊖⊖ **very low**[3,4]	Studies included participants with LCOS or CS due to HF or AMI
	187 per 1000	**90 per 1000** (22 to 363)				
	High[2]					
	500 per 1000	**240 per 1000** (60 to 970)				

[3]Downgraded one step due to study limitation because of lack of blinding of participants and physicians, and missing information on randomisation in the larger study.

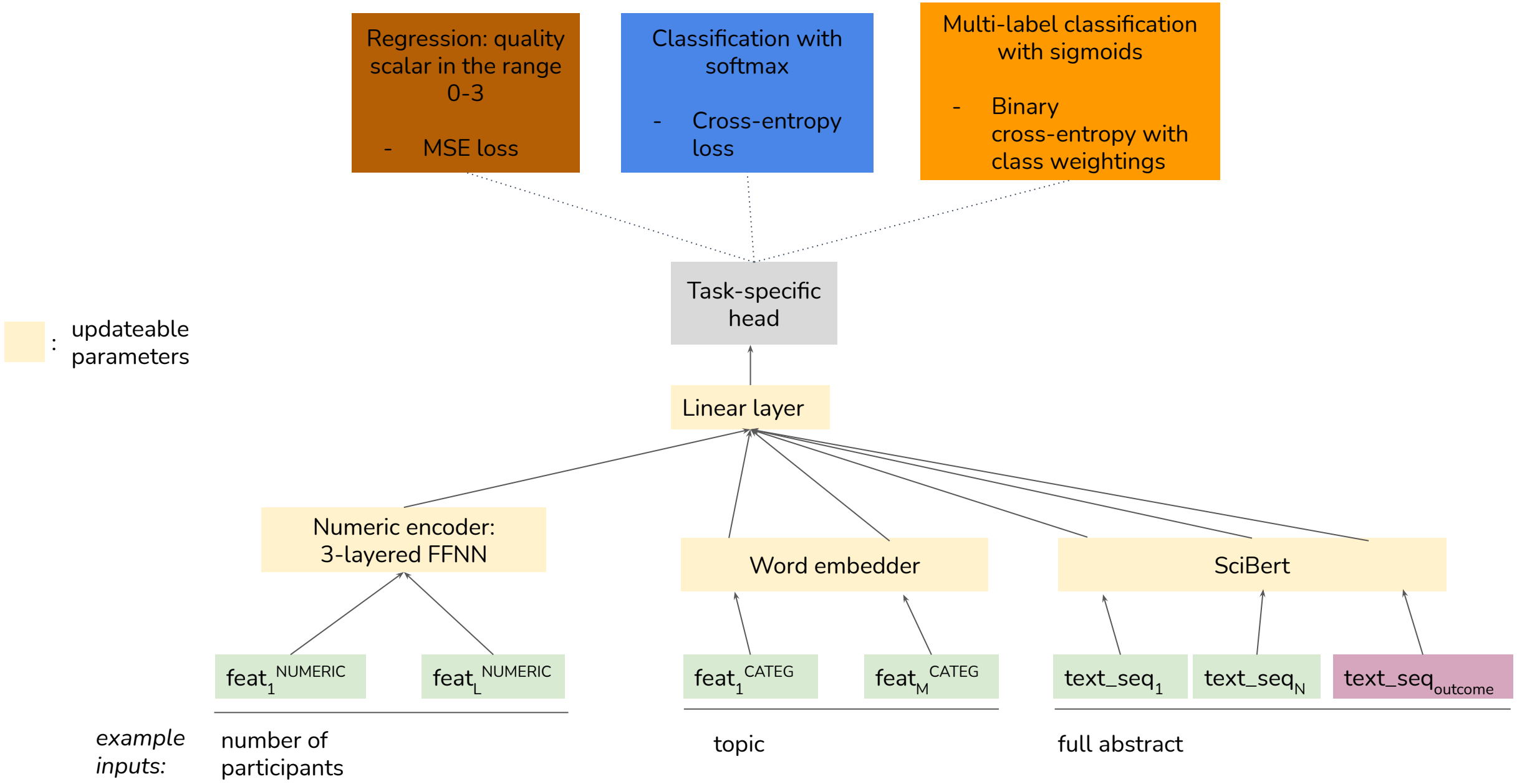
[4]Downgraded two steps for imprecision due to few events and the confidence interval crosses the line of no difference and includes possible benefit from both approaches.

Tasks



classification
regression
multi-label classification

Base model


Regression: quality scalar in the range 0-3
- MSE loss
Classification with softmax
- Cross-entropy loss
Multi-label classification with sigmoids
- Binary cross-entropy with class weightings
Task-specific head
: updateable parameters
Linear layer
Numeric encoder: 3-layered FFNN
Word embedder
SciBert
feat$_1^{NUMERIC}$
feat$_L^{NUMERIC}$
feat$_1^{CATEG}$
feat$_M^{CATEG}$
text_seq$_1$
text_seq$_N$
text_seq$_{outcome}$
example inputs:
number of participants
topic
full abstract

Feature space

Textual

- parts of SRs that are likely to discuss quality
- impose little assumptions, open-ended solution
- (4 in total) *Authors' conclusions*, *plain language summary*, *abstract conclusion*, *full abstract*

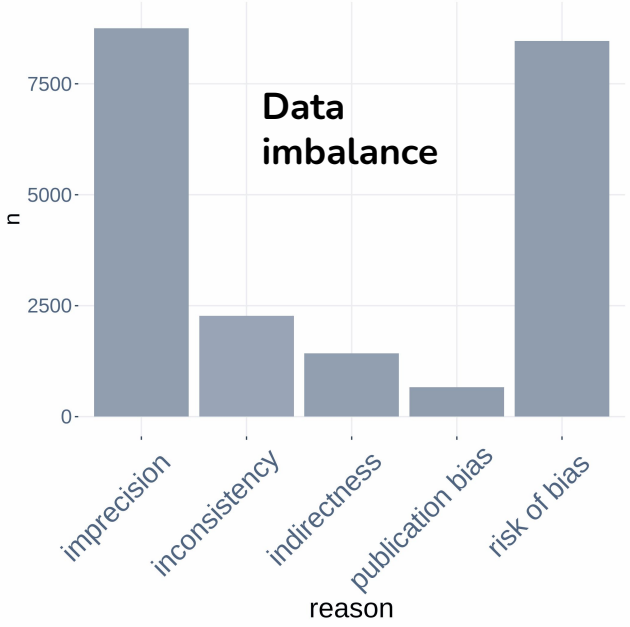
Categorical

- meta-data about the review and non-numerical statistical information
- (3) *Review type*, *topics*, *type of effect*

Numerical

- meta-data about the review and statistics
- (13) *Num. of included studies*, *year*, *num. of outcomes*, *relative effect*, *lower CI*, *upper CI*, ...

Results (averaged over 10 folds)



- classification
- regression
- multi-label classification

Quality score (GRADE)

F1		MAE
0.46	very low 0	
0.51	low 1	
0.53	moderate 2	
0.47	high 3	
0.49		**0.62**
(P: 0.54 R: 0.49)		random: 1.2
random: 0.25		majority: 0.8
majority: 0.13		

Binary quality score

F1	
0.76	very low-low
0.72	moderate-high
0.74	
(P: 0.75 R: 0.75)	
random: 0.43	
majority: 0.46	

Downgrading reasons / **Number of reason types**

F1			MAE
0.63	imprecision	0	
0.21	inconsistency	1	
0.26	indirectness	2	
0.11	pub. bias	3	
0.70	RoB		
0.38			**0.68**
(P: 0.45 R: 0.36)			random: 0.98
random: 0.29			majority: 0.76
majority: 0.29			

Reliability of ratings

Existing user studies (Meader et al., 2014, Berkman et al., 2013, Hartling et al., 2013, Mustafa et al., 2013, Atkins et al., 2005)

- Limited by small sample sizes and datedness
- Poor to almost perfect agreement
- RoB and imprecision (risk of random errors)

What we know so far:

- 4-level quality annotation perhaps too granular/fine distinctions somewhat arbitrary
- Binarisation makes the task easier
- Fairly good performance on some reason classes
  - data imbalance likely a problem (less common reasons predicted less well)
- Our small-scale reliability study for risk-of-bias on recurring primary studies

Possible future work:

- Expert independently grades evidence and provides justifications; then measure agreement with the Cochrane authors →hard and time-consuming.
- Expert only judges the validity of assigned quality grades and justifications; identifies support within the review for authors' decision
- Multiple reviews for the same PICO question -> trouble defining the equivalent questions; unlikely to have the same set of primary studies

Cochrane's RoB2 framework for RCTs

Judgement per type

$judgement_0$
$judgement_1$
...

Overall judgement: low, some concern, high

Risk-of-bias types

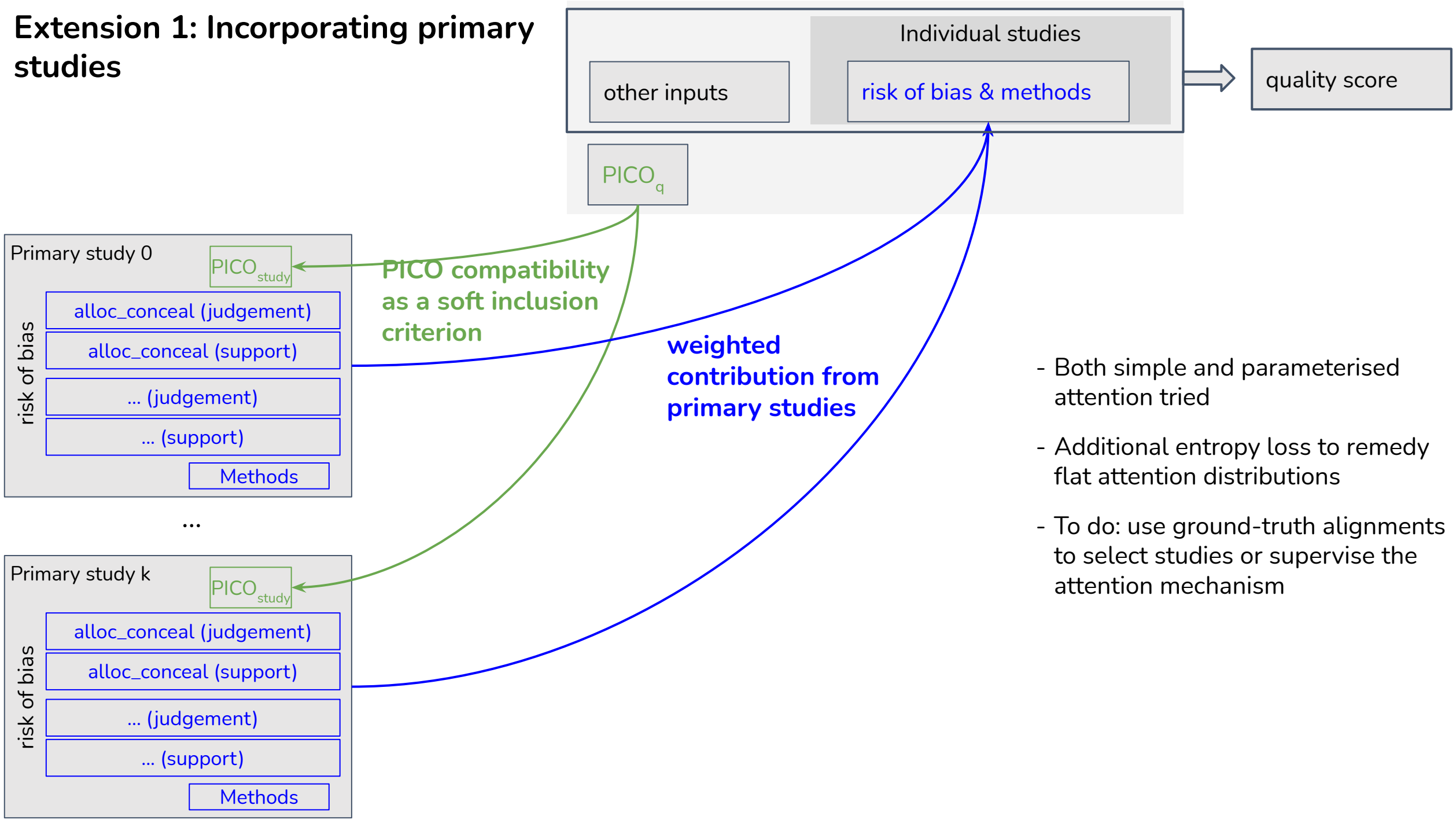
- Bias arising from the randomisation process
- Bias due to deviations from intended interventions
- Bias due to missing outcome data
- Bias in measurement of the outcome
- Bias in selection of the reported result

Signalling questions

- is the allocation sequence random?
- is the allocation sequence adequately concealed?
- do baseline differences between intervention groups suggest a problem with the randomization process?

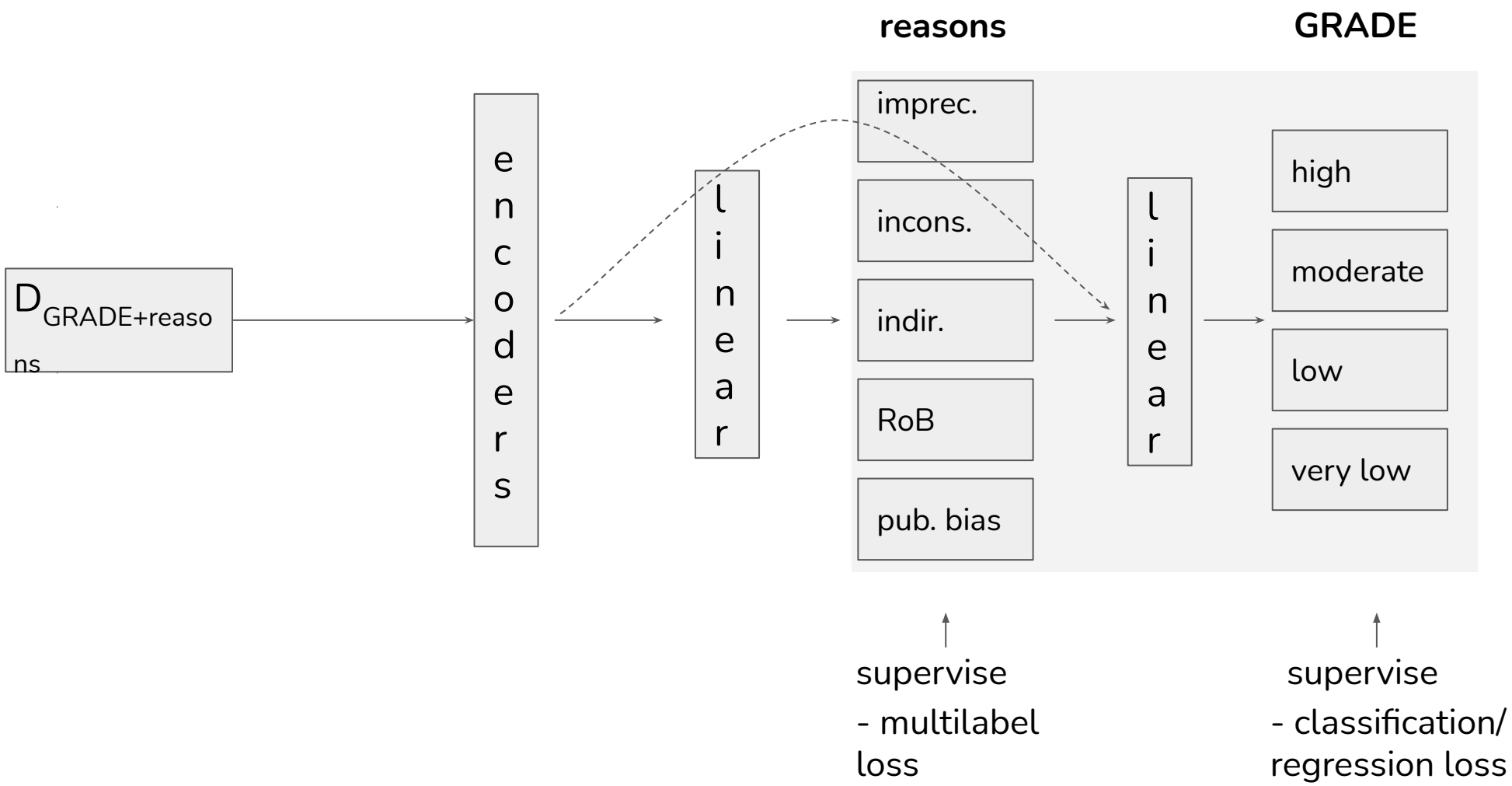
- participants were aware of their assigned intervention during the trial;
- carers and people delivering the interventions were aware of participants' assigned intervention during the trial.
- ...

Extension 1: Incorporating primary studies



- Both simple and parameterised attention tried
- Additional entropy loss to remedy flat attention distributions
- To do: use ground-truth alignments to select studies or supervise the attention mechanism

Extension 2: Joint modeling of downgrading reasons and quality score



Ongoing work and open questions

Obtain data about included studies per PICO
- extract from review editing software files
- observe effect of augmented input on predictions

Augment data for poorly represented reason types

Reliability study

More insights into dataset:
- e.g. are some interventions more likely to yield high-quality evidence? (e.g. pharmacological vs. surgical)
- cluster/label the PICO criteria and relate to the quality of evidence

In adults without cardiovascular disease, does Mediterranean diet (compared to no dietary intervention) help reduce the risk of cardiovascular disease?

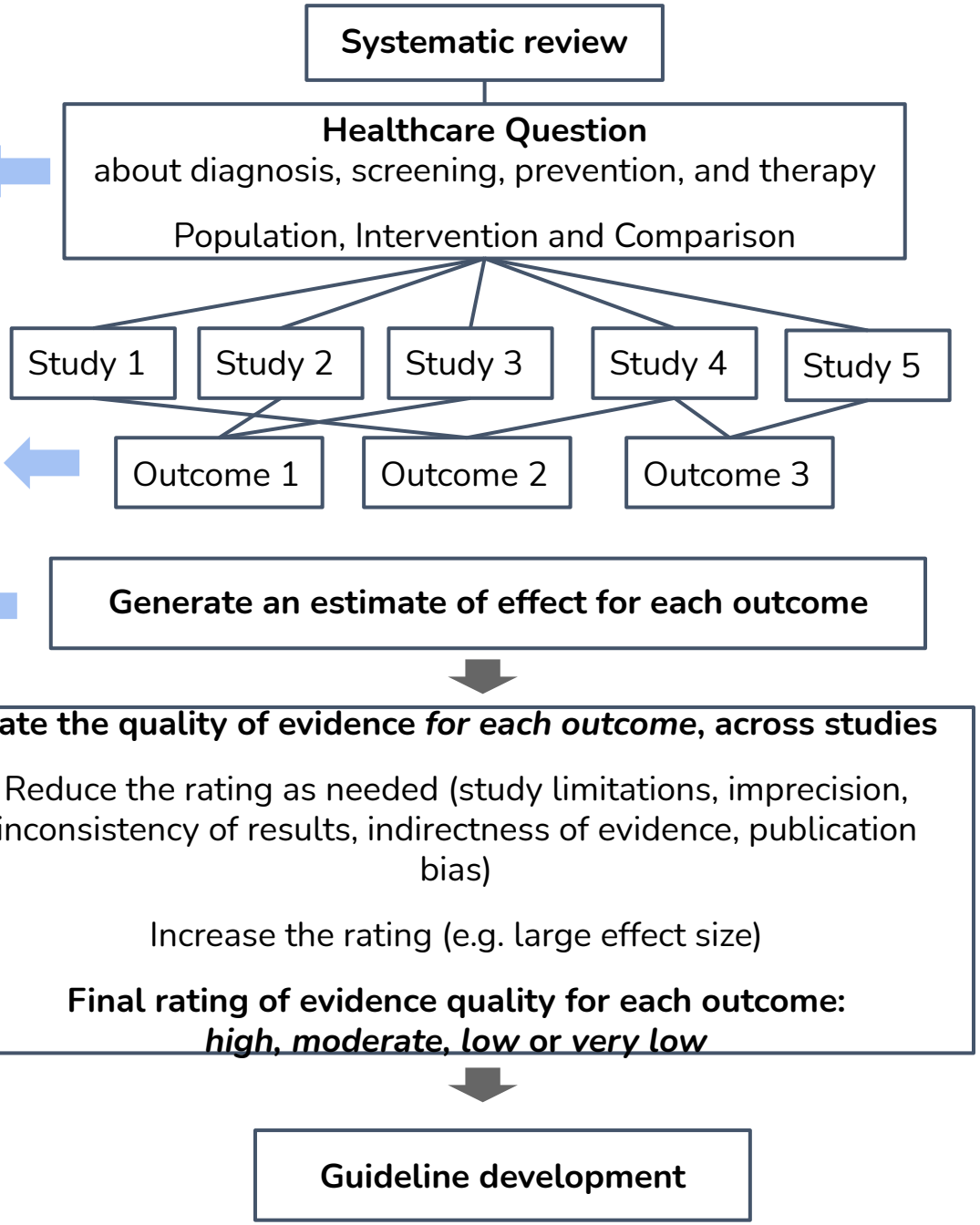
CVD mortality
stroke
myocardial infarction
total cholesterol change
...

Myocardial infarction as outcome:
Risk: 12 per 1000 (Intervention)
16 per 1000 (Control)
...

GRADE: ⊕⊕⊖⊖ (low)

Downgraded by one level for imprecision. Confidence interval is wide enough to include both an important increase or decrease in the outcome.

Downgraded by one level for risk of bias. The only included study was the PREDIMED trial retracted due to methodological issues with randomisation [...]



Modified from Guyatt et al. 2011