



ARC TRAINING CENTRE IN COGNITIVE COMPUTING FOR MEDICAL TECHNOLOGIES

Automated quality assessment of medical evidence



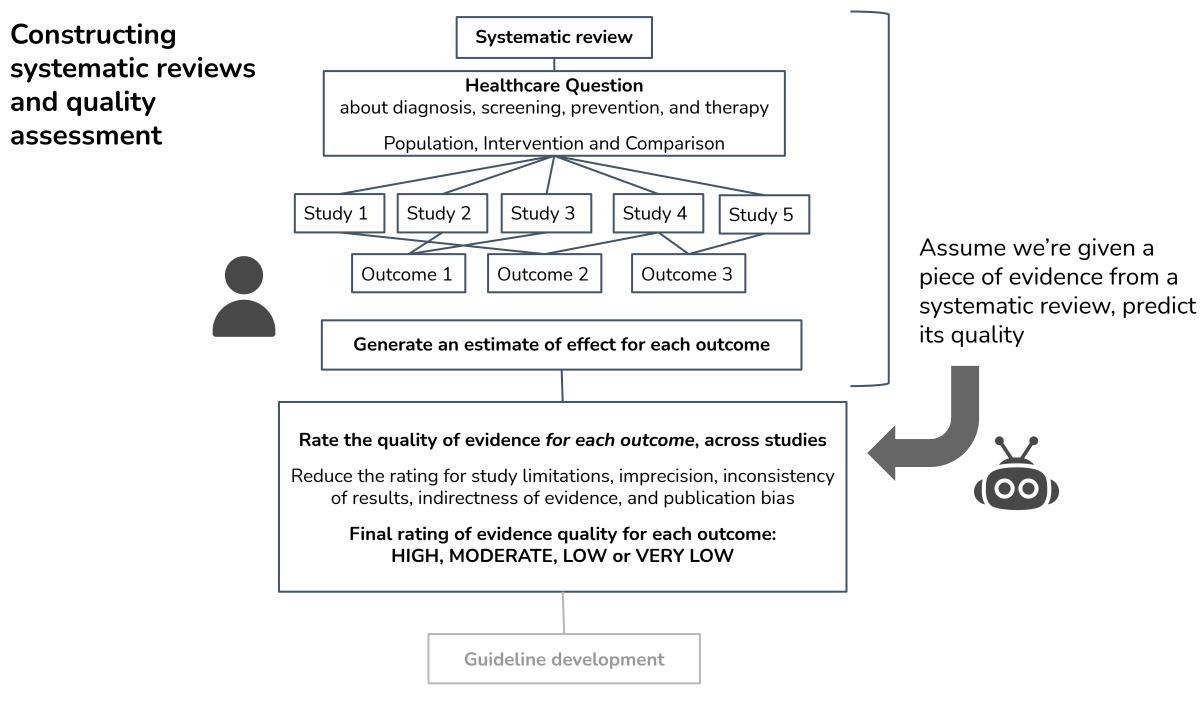


with Tim Baldwin, Antonio Jimeno Yepes, Jey Han Lau, David Martinez Iraola, Yulia Otmakhova, Karin Verspoor

Stream 4

22/3/2021





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



nterventions for re- atheter lumens	Storing paten Contrare Systematic Review - Intervention): Version publicated 29 Databer 2005 https://www.wglite.tot2/14053555.ccm11545.gen210*	ole with
ichrane Systematic Review - Interventin tps://doi.org/33.1002/34653858.CD607		
Thed in 3 publicity	Waala Livingstone Jennifer Hamatty Rupert NcShane Geraldine Hacdonald View article Informati View authors' declarations at interest.	
Clare van Miert Rebecca Hill U		
ew authors' declarations of interest		ppe all Expand al
	Cochange Database of Systematic Reviews Mediterranean-style diet for the primary and secondary prevention of	
	cardiovascular disease	
bstract	Cochrane Systematic Review - Intervention Version published: 13 March 2020 see what's new	
alubirio English Español	https://doi.org/10.3002/34653858.CD009835.pub3 (P	
ackground	New search Coeclusions changed	on. Alzheimer's
-	(n) 100 Yew article information	n with powe
Cochrane Database of Systematic B	🖀 Karen Rees Andrea Takeda Nicole Martin Leila Ellis Dilini Wijesekara Abhinav Vepa Archik Das Louise Hartley	beir effectiveness
Wheat flour fortific	Saverio Stranges	
improving iron stat	View authors' declarations of interest	
Cochrane Systematic Review - Internets Hitps://doc.org/10.1383/34653858.CBD		
Lan	Collapse all Expand all	eating cognitive
	Abstract -	
Martha 5 Field Prasanna Nithra	Available in English Deutsch Español Français 简体中文	
	Background	ve inprovement
	background	cked the references
Abstract	The Seven Countries study in the 1960s showed that populations in the Mediterranean region experienced lower coronary heart	ers to identify
Aunitable in English Español	disease (CHD) mortality probably as a result of different dietary patterns. Later observational studies have confirmed the benefits	
Background	of adherence to a Mediterranean dietary pattern on cardiovascular disease (CVD) risk factors but clinical trial evidence is more limited.	
-		
meet the body's shuriologic seeds	Objectives	es for adults (aged tebo or no
	To determine the effectiveness of a Meditemanean-style diet for the primary and secondary prevention of CVD.	ecourio
t Objectives		
To determine the benefits and har	Search methods	
	We searched the following electronic databases: the Cochrane Central Register of Controlled Trials (CENTRAL; 2018, Issue 9);	
Search methods	MEDLINE (Ovid, 1946 to 25 September 2018); Embase (Ovid, 1980 to 2018 week 39); Neb of Science Core Collection (Thomson	
We searched CENTRAL, MEDLINE, E	Resters, 1900 to 26 September 2018); DARE Issue 2 of 4, 2015 (Cochrane Library); HTA Issue 4 of 4, 2005 (Cochrane Library); NHS	
Selection criteria	EED issue 2 of 4, 2015 (Cochrane Library). We searched trial registers and applied no language restrictions.	
We included cluster- or individual	Selection criteria	
country aged two years and above	the selected randomised controlled trials (HCTs) in healthy adults and adults at high risk of CVD (primary prevention) and those	
other micronutrients. Trials compa	with established CID (secondary prevention). Both of the following key components were required to reach our definition of a	
included.	Mediterranean-style diet: high monounsaturated/saturated fat ratio (use of olive oil as main cooking ingredient and/or	
Data collection and an	consumption of other traditional foods high in monounsaturated fats such as tree nuts) and a high intake of plant-based foods,	
Two review authors independently	including fruits, wegetables and legumes. Additional components included: low to moderate red wine consumption; high	
from included studies and assesses	I risk of blas. We fullowed Cochrane methods in this review.	

systematic reviews (majority from 2010-) Extract data related to quality appraisal from summaries of findings

nma	ary of findings					-
nprovin	ig iron status in population	tified with iron alone compared to unfortified wheat flour (no mic - npared to unfortified wheat flour (no micronutrients added) for re		s added) for re		mia
ing: any		n of all age groups (including pregnant women) from any country over ata for this comparison: Brazil, India, Kuwait, Pakistan, Philippines, Sou on alone				
parisor	Summary of fi	ndings				
omes	Summary of findings for t	he main comparison. Comparison 1: donepezil versus placebo			Open	in table viewer
omia (dr	Patient or population: per Setting: clinic Intervention: donepezil Comparison: placebo					
w WHO sted for opriate	Outcomes	Absolute effects* (95% CI)	Relative effect (95% CI)	Number of participants (studies)	Quality of the evidence (GRADE)	Comments
ths deficie y autho		 The mean change in cognitive abilities in the intervention groups was 0.32 points higher (0.27 lower to 1.31 higher) 		165 (3 RCTs)	e⊕s⊝ Low'	
wup:ra	Behavioural problems (various scales) Follow-up: 12 to 24 weeks	 The mean change in behavioural problems the intervention groups was 0.42 points higher (0.06 lower to 0.89 higher) 		157 (3 RCTs)	eeso Low	
noglob w up: ra ths		Risk Risk with donepezil with placebo	•	•	•	
ity to har about th	Adverse events Follow-up: 12 to 24 weeks	Study population 351 per 630 per 1000 1000 (467 to 768)	OR 0.32 - (0.16 to 0.62)	192 (4 RCTs)	eeso Low ³	
on, alloc		Moderate 157 per 370 per 1000 1000 (222 to 533)				
	Carer stress	No data No data available available	÷	-	10	-
	Institutional/home care	No data No data available available		-	-	

- 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)

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

Settings: hospital

Intervention: levosimendan

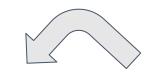
Comparison: placebo

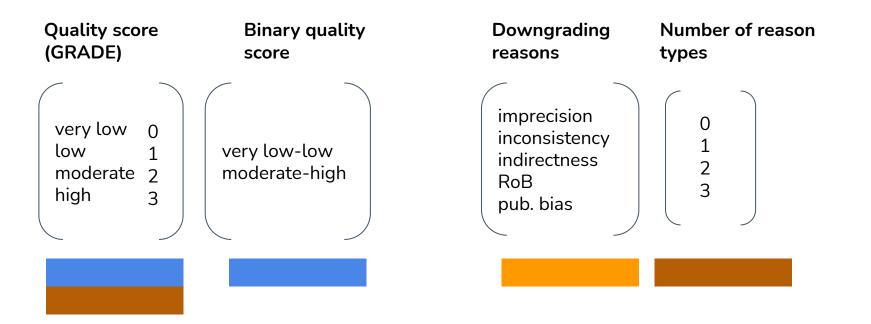
Outcomes	Anticipated (95% CI)	absolute effects	Relative effect (95% CI)	No of participants (studies)	Quality	Comments	
	Risk with placebo	Risk with levosimendan					
All-cause short-term mortality:	Moderate ¹		RR 0.48 (0.12	55	⊕⊕⊝⊝	Studies included participants with	
range <mark>4 t</mark> o 6 months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

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

⁴Downgraded 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.

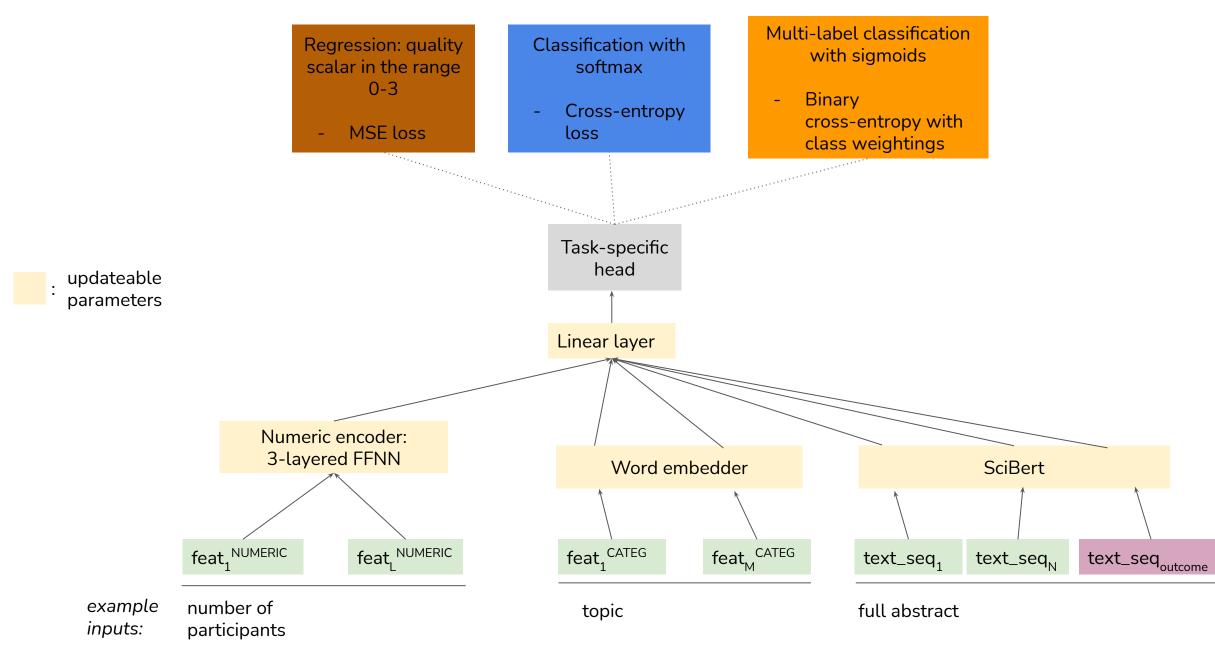
Tasks





classification regression multi-label classification

Base model



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

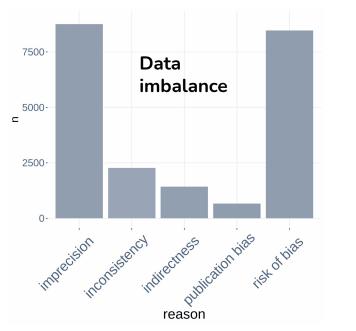
- meta-data about the review and non-numerical statistical information
- (3) Review type, topics, type of effect

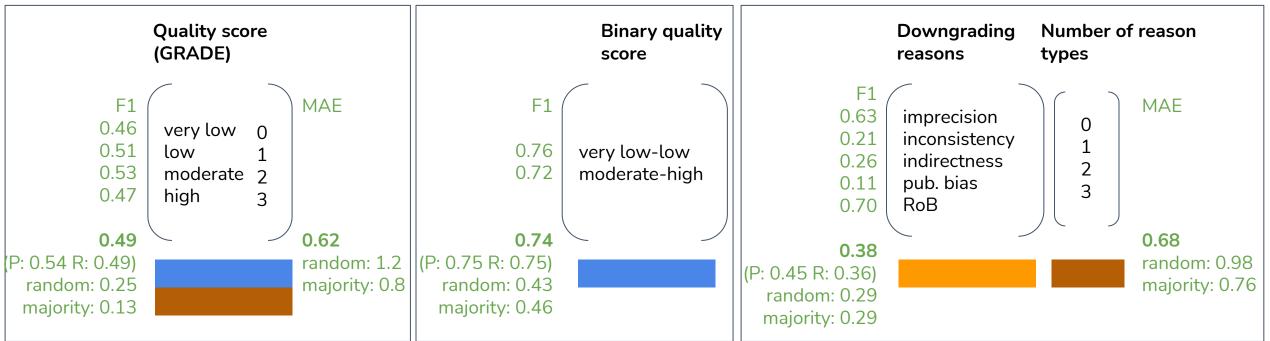
Numerical

- 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





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)

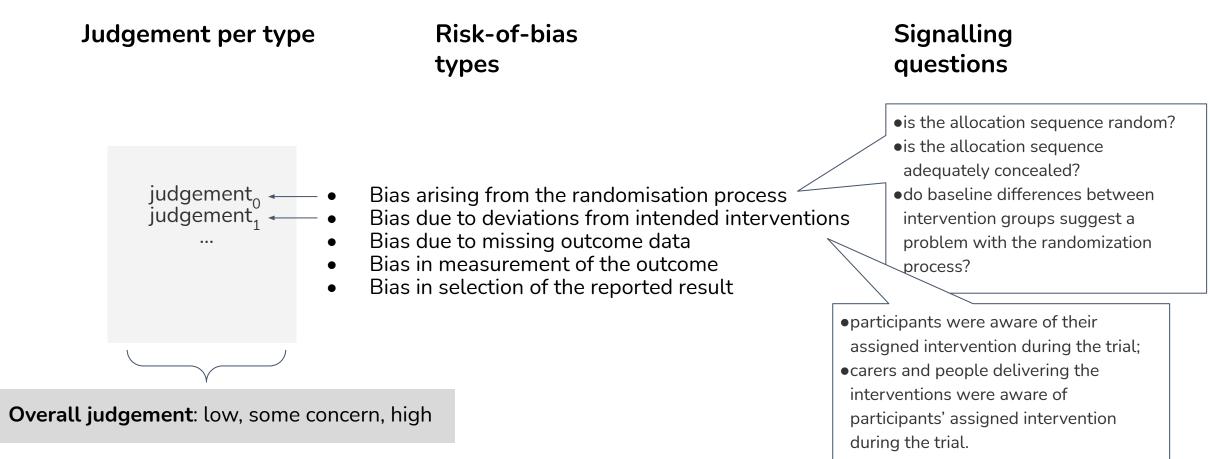
What we know so far:

- 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

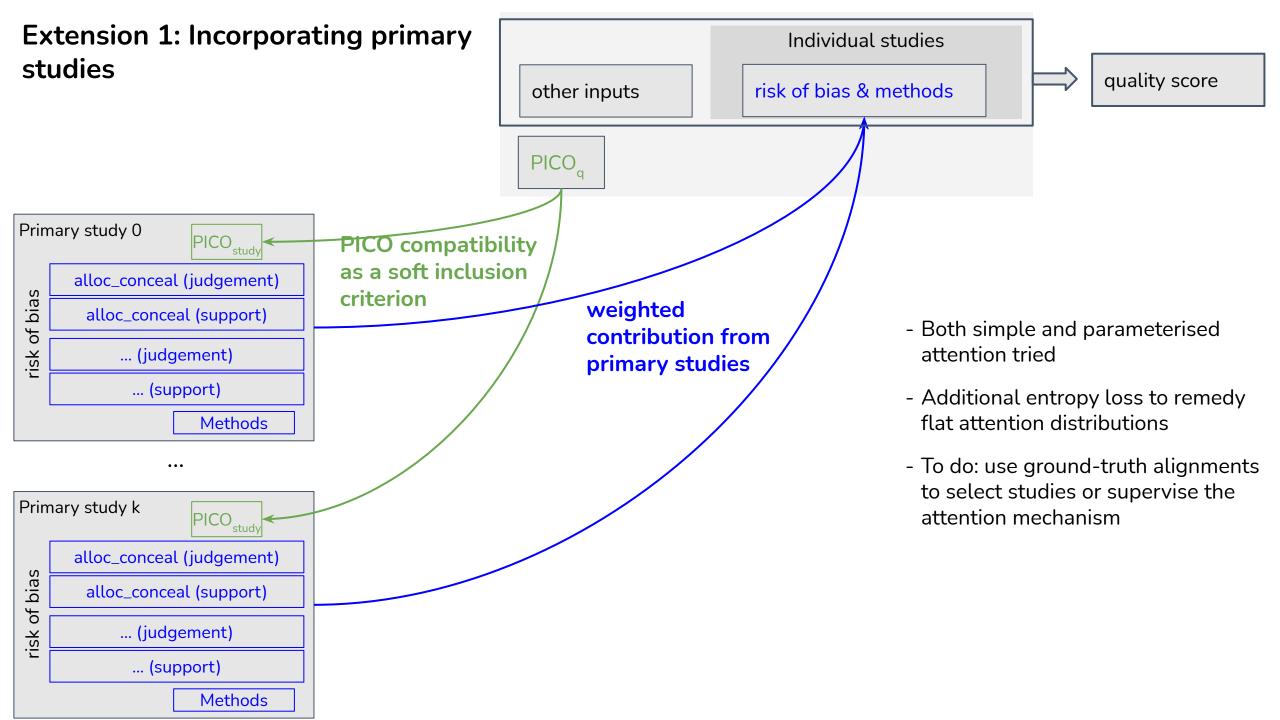
Possible future work:

- Expert independently grades evidence and provides justifications; then measure agreement with the Cochrane authors \rightarrow 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

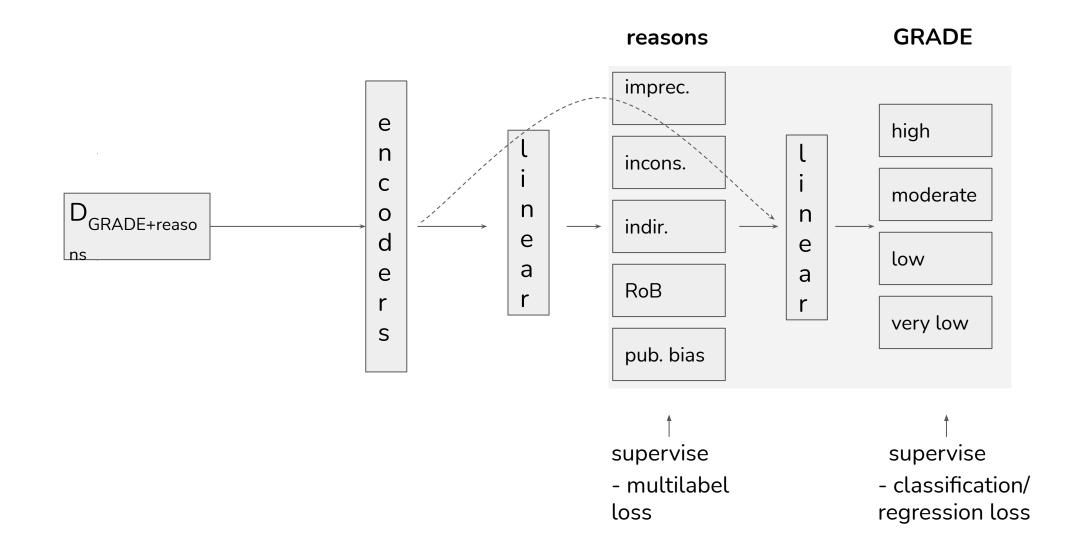
Cochrane's RoB2 framework for RCTs



•...



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

