THE UNIVERSITY OF MELBOURNE

ARC TRAINING CENTRE IN COGNITIVE COMPUTING FOR MEDICAL TECHNOLOGIES

IBM

Automated quality assessment of medical evidence

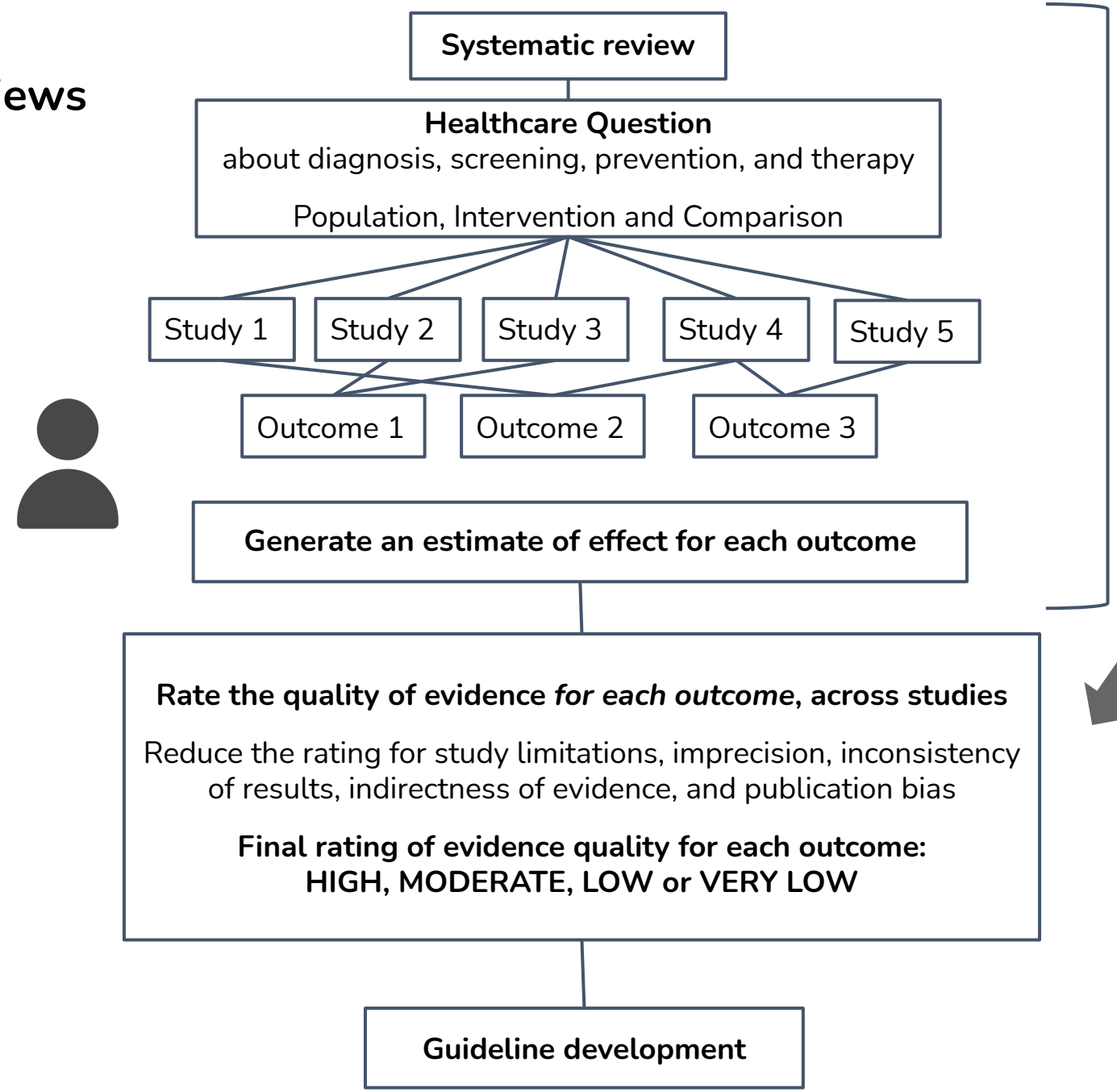
Simon Šuster

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

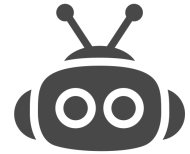
Stream 4

29/1/2021

Constructing systematic reviews and quality assessment



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



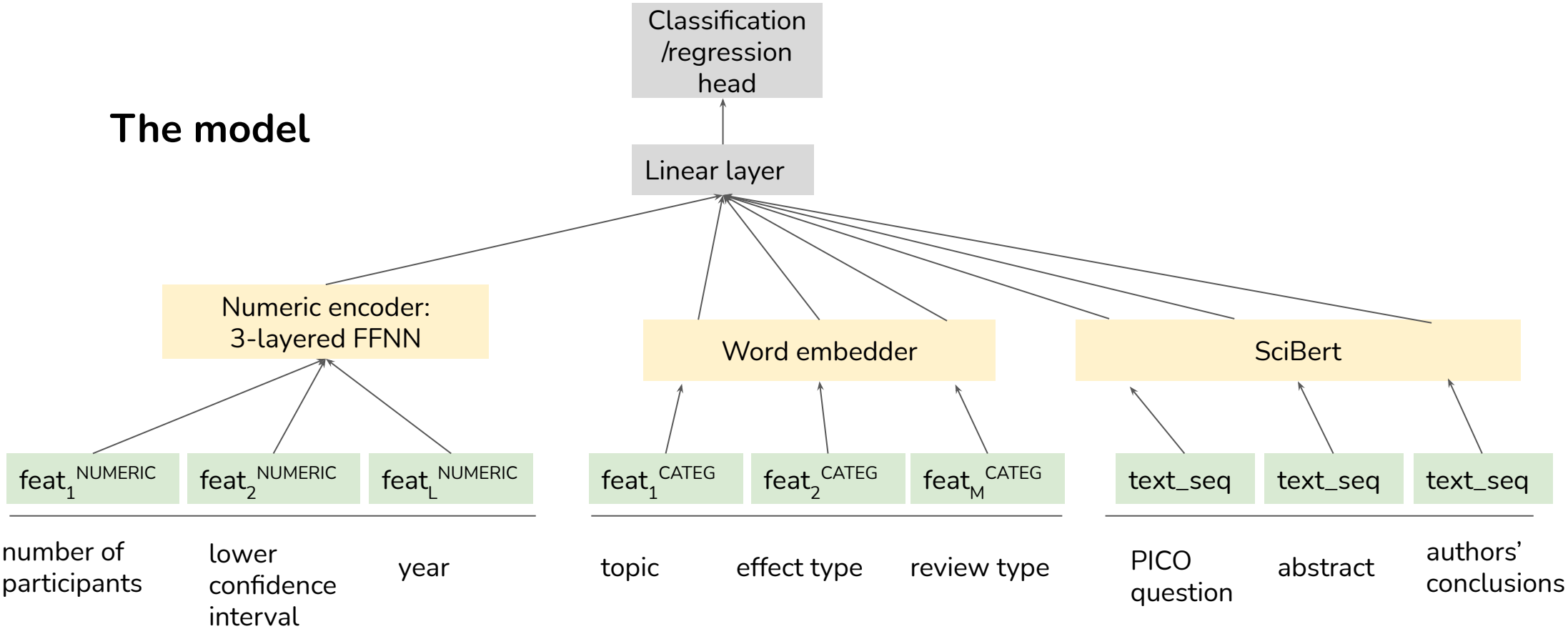
Modified from Guyatt et al. 2011

Obtained from Cochrane Systematic Reviews by extracting relevant data from Summaries of Findings and other parts of reviews.

The data

- ~7,000 reviews
- ~13,500 data points (pieces of evidence), split into train/dev/test sets
- Author-assigned quality scores represent our gold standard (labels)

The model



10-fold cross-validated results

problem:	**4-level grading**		**binarised grading**		**multi-labeling of reasons**	
labels:	high moderate low very low	F1: **0.47** **0.53** **0.51** **0.46**	high/moderate low/very low	F1: **0.72** **0.76**	risk of bias imprecision inconsistency indirectness publication bias	F1: **0.7** **0.6** **0.2** **0.3** **0.1**
overall result:	**F1: 0.5** **MAE: ~0.6** random: 0.25 F1, 1.2 MAE majority: 0.13 F1, 0.8 MAE		**F1: 0.74** random: 0.52 F1 majority: 0.35 F1		**F1: 0.4** random: 0.3 F1 majority: 0.3 F1	

Ongoing work and open questions

Incorporating information from primary studies when assessing overall quality

- problem of retrieval and efficient encoding
- *possible solution*: attentive module that selects the studies based on similarity between questions in the studies and the piece of evidence

Using different label types in a single model

- instances annotated with quality scores *and* downgrading reasons
- *possible solution*: a stacked/multi-task model

If grading reliability of human reviewers is poor, how does it affect learning?

- *possible solution*: obtain multiply graded pieces of evidence and report interrater agreement; are there categories with higher reliability?

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

Systematic review

Healthcare Question
about diagnosis, screening, prevention, and therapy

Population, Intervention and Comparison

CVD mortality
stroke
myocardial infarction
total cholesterol change
...

Study 1 | Study 2 | Study 3 | Study 4 | Study 5

Outcome 1 | Outcome 2 | Outcome 3

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

Generate an estimate of effect for each outcome

Rate the quality of evidence *for each outcome*, across studies

Reduce the rating as needed (study limitations, imprecision, inconsistency of results, indirectness of evidence, publication bias)

Increase the rating (e.g. large effect size)

Final rating of evidence quality for each outcome:
high, moderate, low* or *very low

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 [...]

Guideline development

Modified from Guyatt et al. 2011