UNIVERSITY
OF OSLO

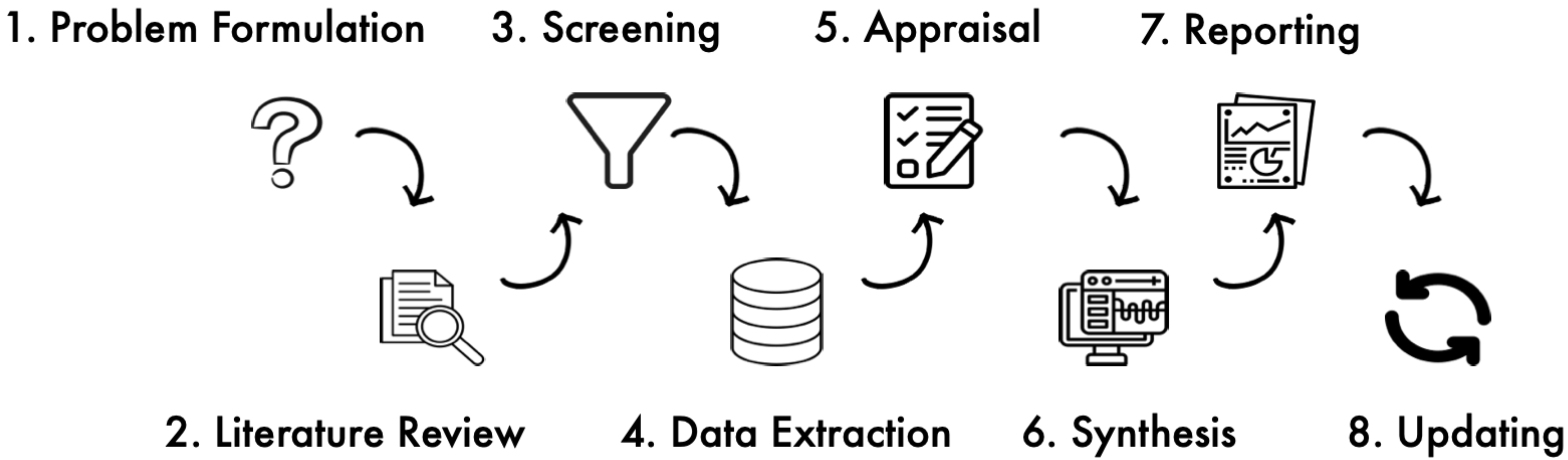
AIRES: Accelerating Research Synthesis in Educational Sciences

Diego G. Campos

Centre for Educational Measurement
Centre for Research on Equality in Education
University of Oslo, Norway

26.02.25

Steps in a Systematic Reviews and Meta-Analysis



AI Tools for Systematic Literature Reviews and Meta-Analyses in Educational Psychology: A Comprehensive Review and a Guide

Tim Fütterer[1], Diego G. Campos[2,3], Thomas Gfrörer[1], Rosa Lavelle-Hill[1,4], Kou Murayama[1], and Ronny Scherer[2,3]

Received: 24 January 2024 | Revised: 24 July 2024 | Accepted: 4 September 2024
DOI: 10.1002/jrsm.1762

RESEARCH ARTICLE

Research Synthesis Methods WILEY

An evaluation of the performance of stopping rules in AI-aided screening for psychological meta-analytical research

Lars König[1] | Steffen Zitzmann[2] | Tim Fütterer[3] | Diego G. Campos[4] | Ronny Scherer[4] | Martin Hecht[1]

UNIVERSITY OF OSLO

Educational Psychology Review (2024) 36:19
https://doi.org/10.1007/s10648-024-09862-5

REVIEW ARTICLE

Screening Smarter, Not Harder: A Comparative Analysis of Machine Learning Screening Algorithms and Heuristic Stopping Criteria for Systematic Reviews in Educational Research

Diego G. Campos[1] · Tim Fütterer[2] · Thomas Gfrörer[2] · Rosa Lavelle-Hill[2,3] · Kou Murayama[2] · Lars König[4] · Martin Hecht[4] · Steffen Zitzmann[2] · Ronny Scherer[1]

Accepted: 29 January 2024 / Published online: 8 February 2024

Workshop
AI-Tools for Systematic Reviews and Meta-Analyses

Dr. Tim Fütterer, Dr. Thomas Gfrörer, Dr. Diego G. Campos
Workshop | 30.07.2024 | online

HECTOR-INSTITUT FÜR EMPIRISCHE BILDUNGSFORSCHUNG

AI Tools for Systematic Literature Reviews and Meta-Analyses in Educational Psychology: A Comprehensive Review and a Guide

Tim Fütterer[1], Diego G. Campos[2,3], Thomas Gfrörer[1], Rosa Lavelle-Hill[1,4], Kou Murayama[1], and Ronny Scherer[2,3]

(RQ1) What AI tools are available for each step in the systematic review and meta-analysis process implemented in educational psychology?

(RQ2) What are AI tools' capabilities, limitations, and domain-specific challenges for conducting systematic reviews and meta-analyses in educational psychology?

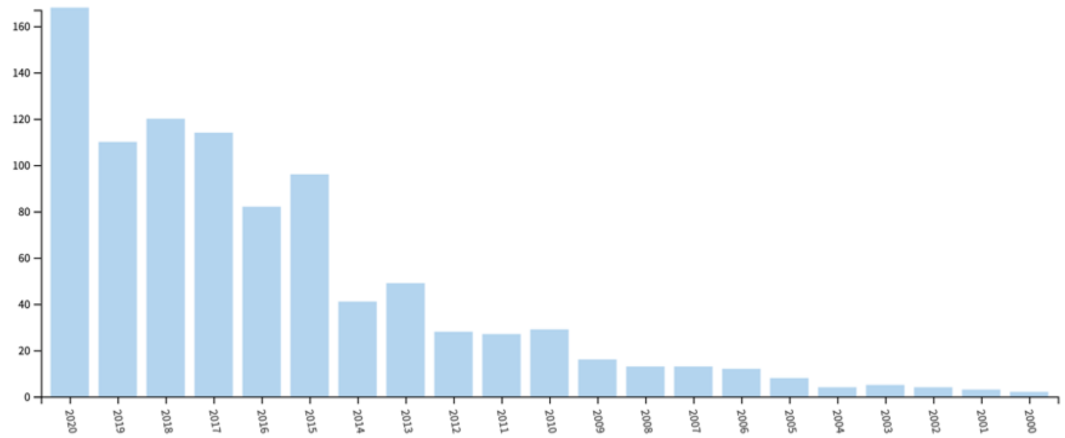
(RQ3) How can researchers use AI tools to improve the efficiency, accuracy, and scalability of systematic reviews and meta-analyses?

(RQ4) How do researchers (in educational psychology) perceive the integration of AI tools into systematic reviews and meta-analyses?

Nr.	Inclusion criteria	Exclusion criteria
1	Artificial Intelligence (AI) used Tools to be included must use any form of AI method. Keywords that can indicate AI are, for instance, Machine Learning, Artificial Intelligence, Classification, Classifier, Clustering, Deep Learning, Reinforcement Learning, Supervised Learning, Unsupervised Learning, Cross Validation, Computer Vision, Natural Language Processing (NLP), Random Forest, Decision Trees, XG Boost, Ada Boost, Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbor, Multi-Layer Perceptron (MLP), Artificial Neural Network (ANN), Deep Neural Network (DNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long-Term Short-Term Memory (LSTM), Hidden Markov Models (HMM), K-Means, Gaussian Mixture Models, Prediction, Accuracy, F1 Score, AUROC (Area Under the ROC)	No AI used Tools that do not use any form of AI method were excluded (e.g., tools such as the reference management software Zotero)
2	Data security / protection guaranteed (e.g., user data are stored locally, personal data are not collected)	Data security / protection not guaranteed (e.g., user data are stored on a server, personal data are collected)
3	Transparency of technology used (e.g., descriptions of the ML -algorithms used are available)	No transparency of technology used (e.g., description of the ML -algorithms used are not available)
4	Tool (freely) available (e.g., an online link / access to the tool is available, the tool is being patched / updated)	Tool not (freely) available (e.g., an online link/ access to the tool is not available, the tool is not being patched/ updated at all, you must pay for the tool)
5	Usable in field of educational psychology (EP) (e.g., model trained on or applicable to educational and psychological data / research [i.e., "trained on" meaning that it could be trained on different data but transfers well to the education setting])	Not usable in field of educational psychology (EP) (e.g., model trained on or applicable only to biological data/ research)
6	Installation process easy (e.g., approximately less than 10 min are needed to get the tool running)	Installation process not easy (more than 10 min to get the tool running)
7	Specific tool Tools to be included had to be developed to conduct research syntheses.	Generic tool Tools that are designed to provide, for example, algorithms that can be used for different purposes for a specific use case such as a research thesis, but that were not specifically developed for the implementation of research theses (e.g., NLP approaches) were excluded.

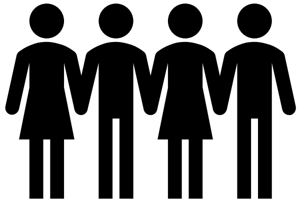
Name	Summary	Step
ASReview	ASReview LAB is a free, open-source machine learning tool that accelerates text screening through a user-friendly interface, extensive simulations to measure efficiency gains, and local installation to ensure data privacy.	2
Rayyan	Rayyan is an AI-driven platform that simplifies and speeds up systematic reviews while enabling free, flexible collaboration of any size across the web or its dedicated app.	2
Trial-streamer	Trialstreamer continuously monitors major trial databases, uses machine learning and rule-based methods to extract and normalize RCT data (including PICO elements, MeSH terms, sample sizes, risk of bias, and key findings) and provides free access to a ranked, searchable database via a web portal.	2, 3, 5
Colandr	Colandr is an open-access, machine-learning driven platform that leverages NLP and text-mining to automate evidence synthesis by identifying relevant citations and extracting data from PDF articles.	2, 3
Abstrackr	Abstrackr is an online platform designed to streamline the citation screening process for systematic reviews.	2
Meta-CART	Meta-CART uses classification and regression trees (CART) to detect interactions between moderators in meta-analysis, followed by subgroup analyses to test the significance of these moderator effects.	5
SWIFT-Review	SWIFT-Review, short for the Sciome Workbench for Interactive computer-Facilitated Text-mining, is a free, interactive platform offering various tools for problem formulation and literature prioritization.	1, 2, 3, 4

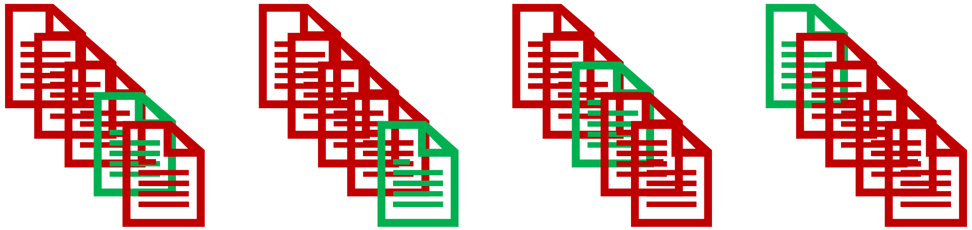
Screening - Traditional



Educational video games 2000 – 2020
Web of Science

The number of publications have exponentially increased





Reviewers screen all documents

5% relevant documents

- Screening fatigue
- 13% missed documents (Waffenschmidt et al., 2019)

UNIVERSITY
OF OSLO

Screening – Utilizing AI

- AI tools can help researchers to find relevant records by using **screening prioritization** with **active learning**

Screening prioritization re-arranges the records to be screened from random to a more intelligent order

Reviewers do not have to screen all documents.

Coding – Traditional

Collect information from relevant records

- Study characteristics (e.g., quality)
- Sample characteristics
- Outcome characteristics
- Effect sizes

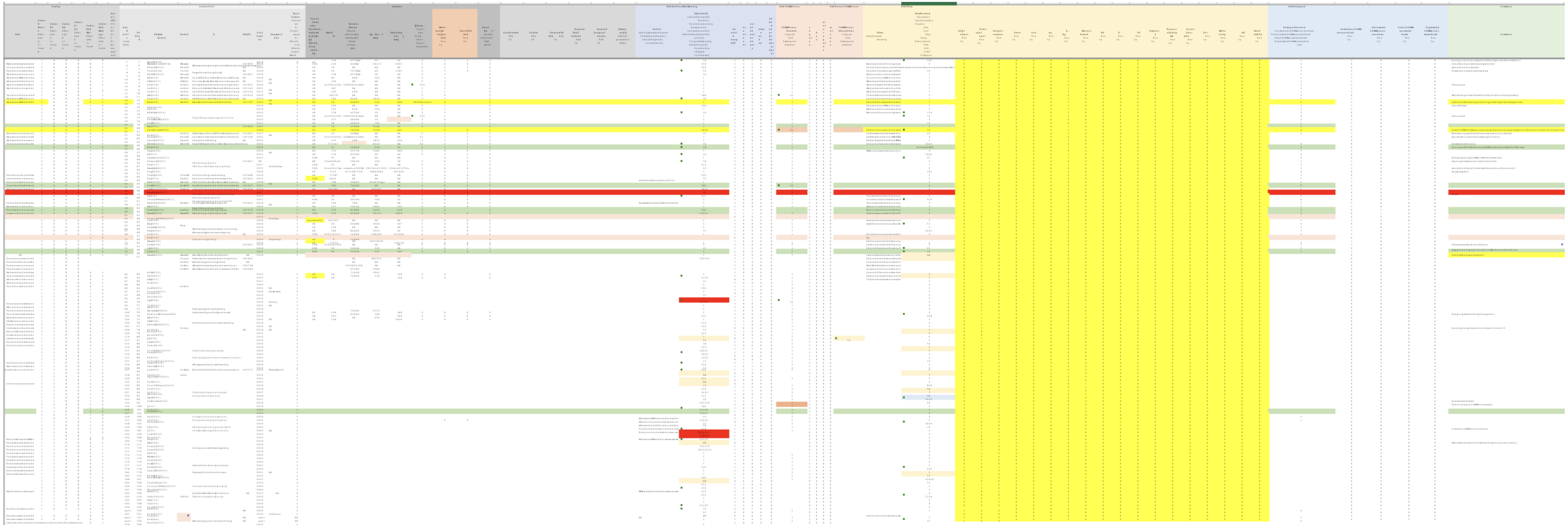
StudyID	Coder	Title	Publication type
S046	JPustejovsky	Efficacy of outpatient aftercare for adolescents with alcohol use disorders: A randomized controlled study.	Journal article
			Technical report
			Journal article

StudyID	SampleID	Coder	Sample description	Percent male	Percent male source	Average age	Average age source
S046	S046-1	DKaplan	Full sample	68	307	17.2	307
S048	S048-1	DKaplan	Male subsample	100	1308	16.4	1308
						17.4	1308
						15.7	72

SampleID	OutcomeID	Coder	Outcome scale	Follow-up time (weeks)	Effect size source
S046-1	S046-1-01	MChen	Alcohol use	8	Table 3
S046-1	S046-1-02	MChen	Alcohol use	12	Table 3
				12	Table 3
				14	Table 7
				14	Table 7

sampleID	Outcome	Follow-up time	ES estimate	ES std. error
S01-1	BDI	12	-0.30	0.22
S01-2	BDI	12	0.04	0.22
S02-1	CES-D	14	-0.09	0.18
S03-1	HAM-D	10	0.03	0.16

Coding – Traditional

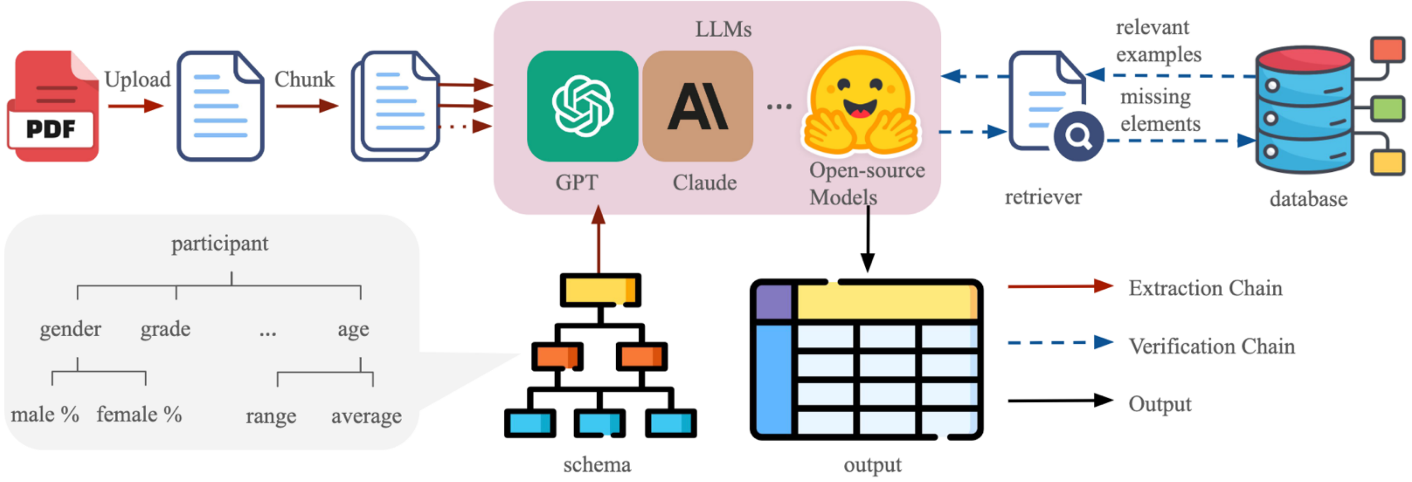


Coding – Utilizing AI

AI tools for data extraction use ML and NLP for text classification to **categorize sentences within documents into predefined categories relevant to the review**

Information extraction attempts to extract the text snippets corresponding to the desired data elements

Researchers provide examples of categories to train the model



Wang, et al., (2024)

UNIVERSITY OF OSLO

Screening Smarter, Not Harder: A Comparative Analysis of Machine Learning Screening Algorithms and Heuristic Stopping Criteria for Systematic Reviews in Educational Research

Diego G. Campos; Tim Fütterer[2], Thomas Gfrörer[2], Rosa Lavelle-Hill[2], Kou Murayama[2], Lars König[3], Martin Hecht[3], Steffen Zitzmann[2], and Ronny Scherer[1]

[1] University of Oslo, Norway; [2] University of Tübingen, Germany;

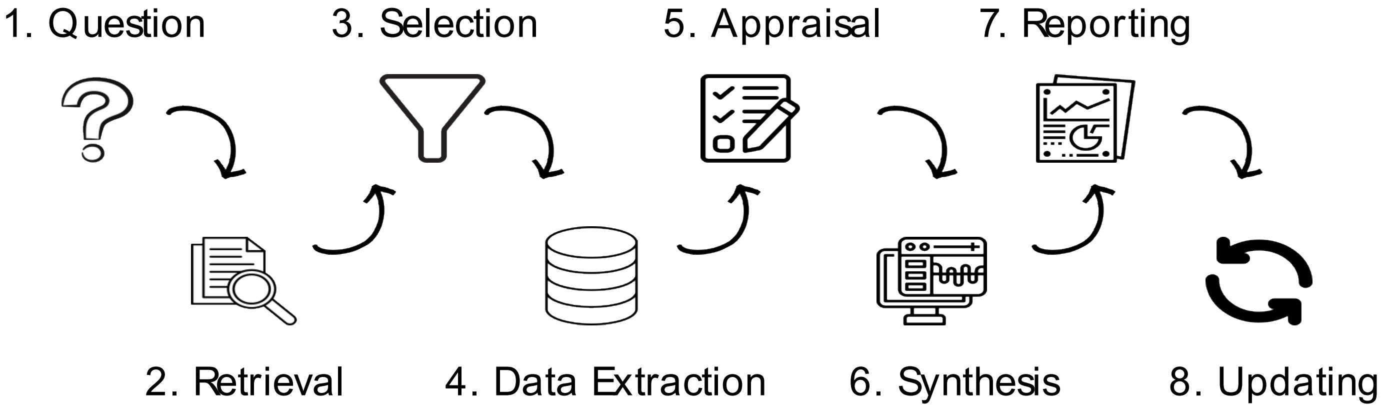
[3]Helmut-Schmidt-Universität, Germany

26.02.25

Introduction

- Systematic reviews and meta-analyses are critical for advancing practice, research, and policy (e.g., Schneider & Preckel, 2017; van de Schoot et al., 2021; van Huizen & Plantenga, 2018).
- Understand teaching and learning processes and to derive measures for an effective educational system (Taylor & Hedges, 2023).
- What works clearinghouses (Hedges, 2018).

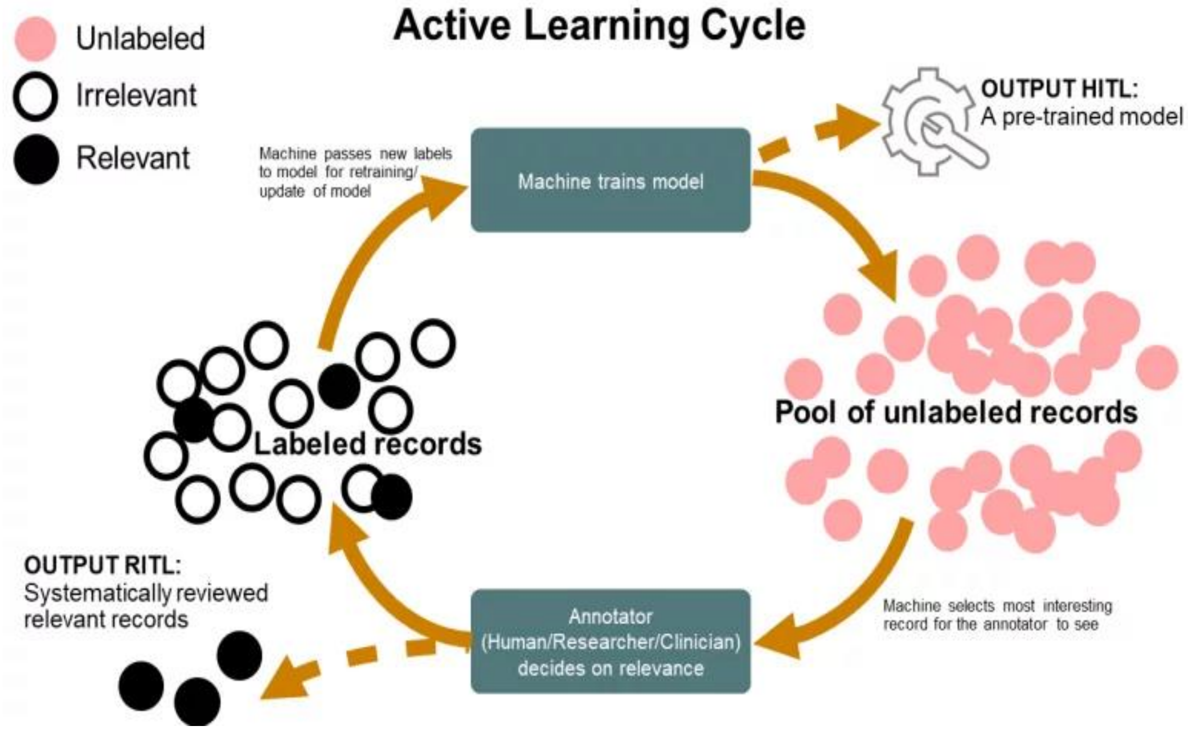
Systematic Review



- John Hattie and his team needed 15 years to complete their meta-analysis “visible learning” (Hattie, 2009).
- An experienced reviewer can take between 30 seconds and 1 minute per reference (Przybyła, et al. 2017).

Machine Learning and Literature Screening

- ASReview can help researchers to find relevant records by using **screening prioritization** with **active learning**



https://asreview.nl/blog/active-learning-explained/

Screening prioritization re-arranges the records to be screened from random to a more intelligent order

Integrating AI tools into the Systematic Review Process

- Simulation studies in health sciences have shown that the AI screening process within ASReview is reliable and can save more than 90% of the time compared to purely human-based screening (van de Schoot et al., 2021).

- The performance of AI machines varies across research domains (van de Schoot et al., 2021).

- Clear stopping rules need to be defined (Yu & Menzies, 2019).

Stopping rules

Statistic Approaches

The decision to stop screening is based on deriving an estimate of the total number of relevant papers within the initial training set (e.g.; van Haastrecht et al., 2021).

Heuristic Approaches

- Time-based (Wallace et al., 2010).
- Data-driven (Ros et al., 2017).
- Mixed-data (Hamel et al., 2021).

Research questions

How do AI-based NLP algorithms perform with respect to sensitivity and ETS on educational data sets?

How do heuristic stopping criteria perform in terms of sensitivity, specificity and ETS on educational data sets?

Methodology and data sources

The study used a systematic approach to compare the accuracy, speed, and cost-effectiveness of learning algorithms for abstract screening in education. Data were collected from multiple academic databases and analyzed using statistical software.

Sample

- Journals in education and educational psychology.
  - Campbell Reviews (n = 17).
  - Review of Educational Research (n = 95).
  - Educational Research Review (n = 83).
  - Educational Psychology Review (n = 37).
- Experts in the field (n = 84).
- Total records sought for retrieval (n = 316).
- Reports excluded (n = 12).
- Total reports included (n = 28).

Sample

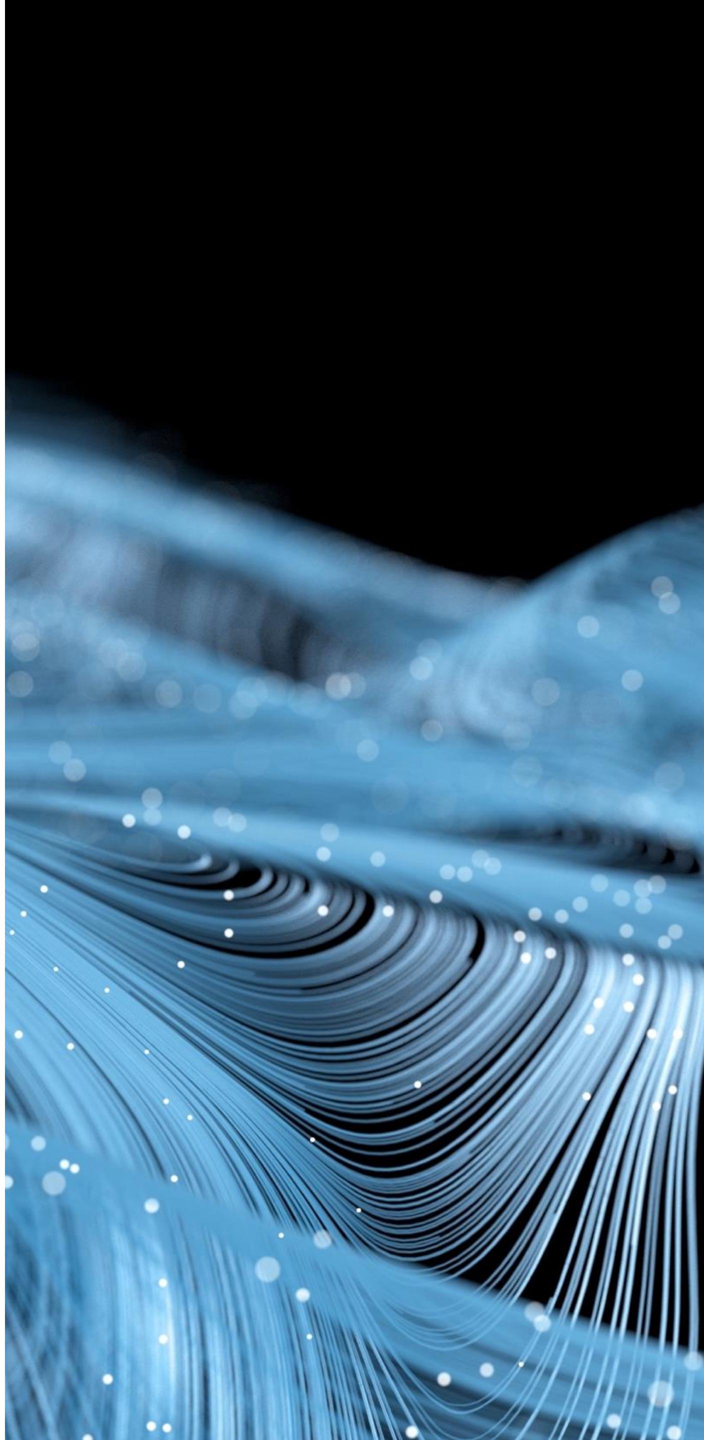
Descriptive statistics of databases included in the simulation study.

Database	Total Refences	References Selected for Full Text Screening	Percentage of References Selected for Full Text Screening
Anmarkrud, et al., (2022)	2550	230	8.5 %
Backfisch, et al., (2020)	2353	719	31.2 %
Capparozza, et al., (2021)	2739	171	6.5 %
Endedijk, et al., (2022)	4452	365	8.5 %
Filges, et al., (2018)	8065	374	4.1 %
Filges, et al., (2022)	16221	6429	9.8 %
Fong, et al., (2021)	2471	199	8.7 %
Fütterer, et al., (2023)	19497	44	0.2 %
Jäger-Dengler-Harles, et al., (2020)	6798	1212	20.2 %
Kupers, et al., (2019)	3094	962	31.1%
Lesperance, et al., (2022)	8465	149	1.8 %
Neri & Retelsdorf, (2022)	686	115	16.8 %
Noetel, et al., (2022)	1368	86	5.8 %
Pico & Woods, (2023)	30	14	46.7 %
Roberts, et al., (2022)	5065	168	3.3 %
Rowan, et al., (2021)	416	209	50.1 %
Saqr, et al., (2022)	4581	1818	39.5 %

Method

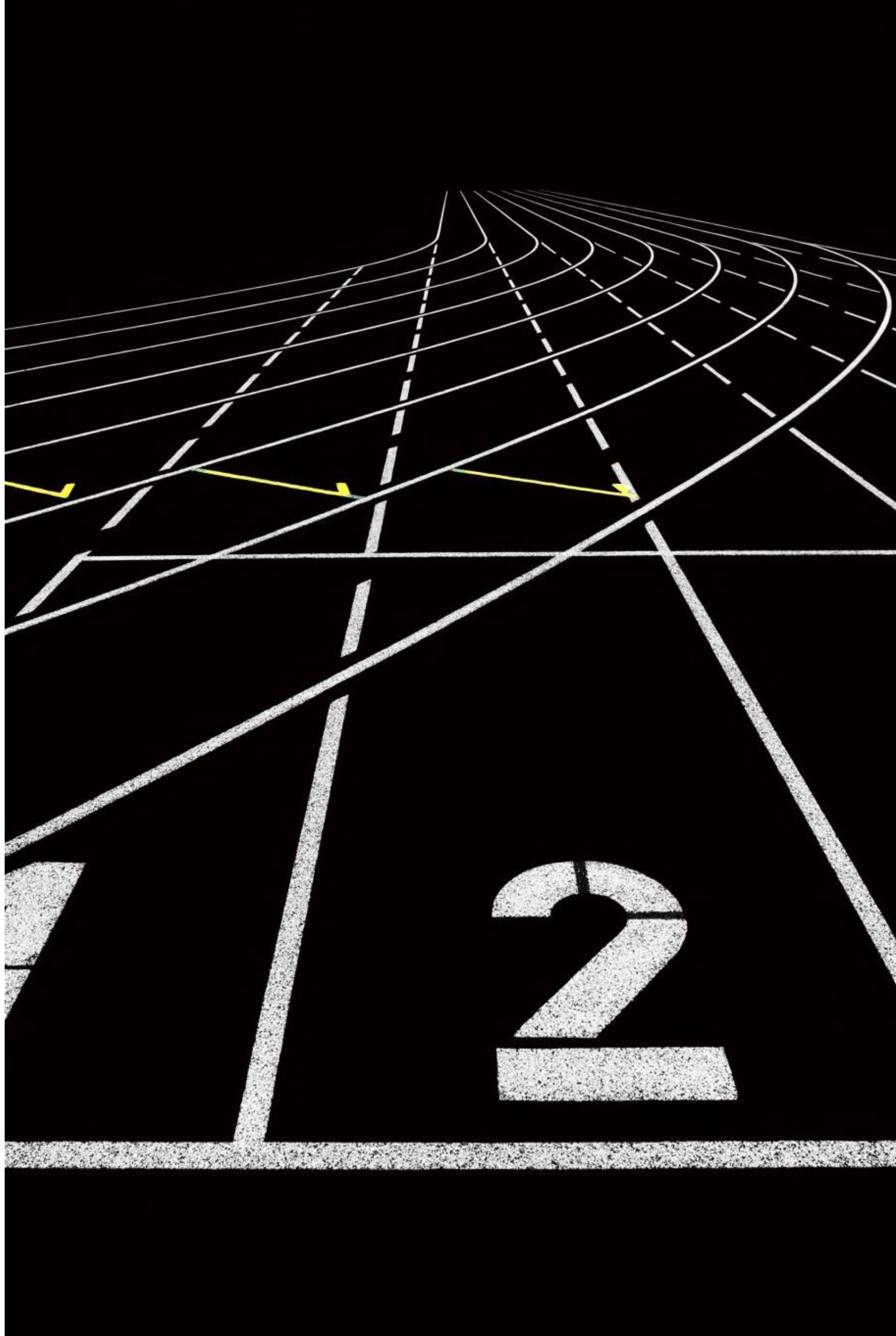
- ASReview (van de Schoot et al., 2021)
- R (R Core Team (2022).

Feature Extraction Technique	**Classifier**	**Learning Algorithm**
Doc2Vec	Naïve Bayes (NB)	LR + Doc2Vec
Sentence BERT (SBERT)	Logistic Regression (LR)	LR + SBERT
Term Frequency-Inverse Abstract Frequency (TF-IDF)	Random Forest (RF)	LR + TFIDF
	Support Vector Machine (SVM)	NB + TF-IDF
		RF + Doc2Vec
		RF + SBERT
		RF + TF-IDF
		SVM + Doc2Vec
		SVM + SBERT
		SVM + TF-IDF



Model Evaluation

- **Sensitivity**:
  - The proportion of relevant studies in an abstract collection identified by the learning algorithms out of the total deemed relevant by the authors of the systematic review
- **Specificity:**
  - The proportion of studies in an abstract collection that would not require screening by reviewers when using learning algorithms out of the total deemed irrelevant by the authors of the systematic review
- **Estimated time savings (ETS):**
  - The reduction in working days achieved when using learning algorithms by not having to screen irrelevant abstracts



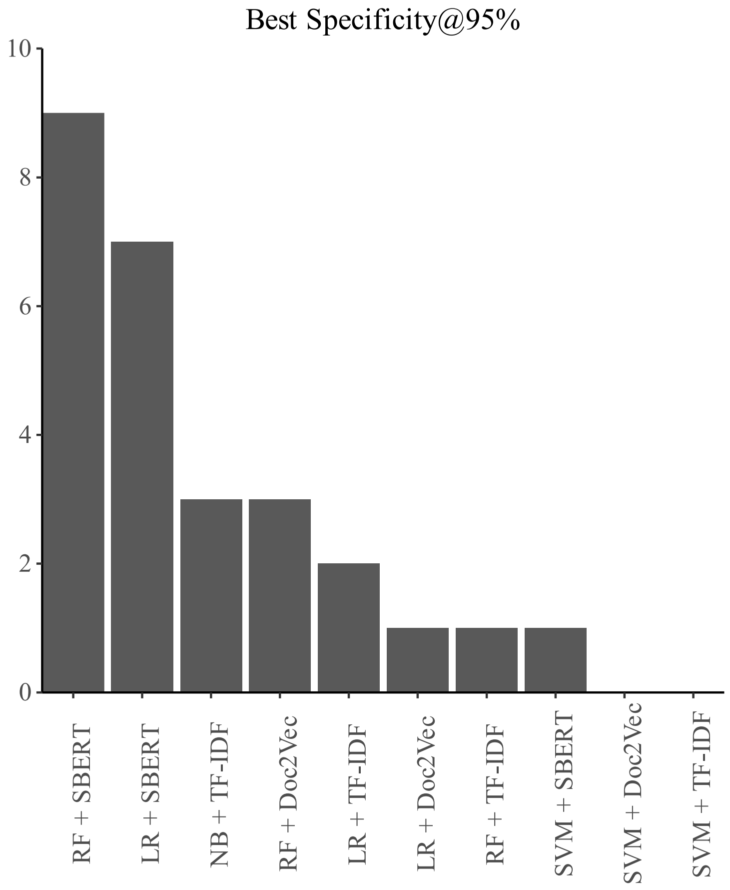
Results and analysis

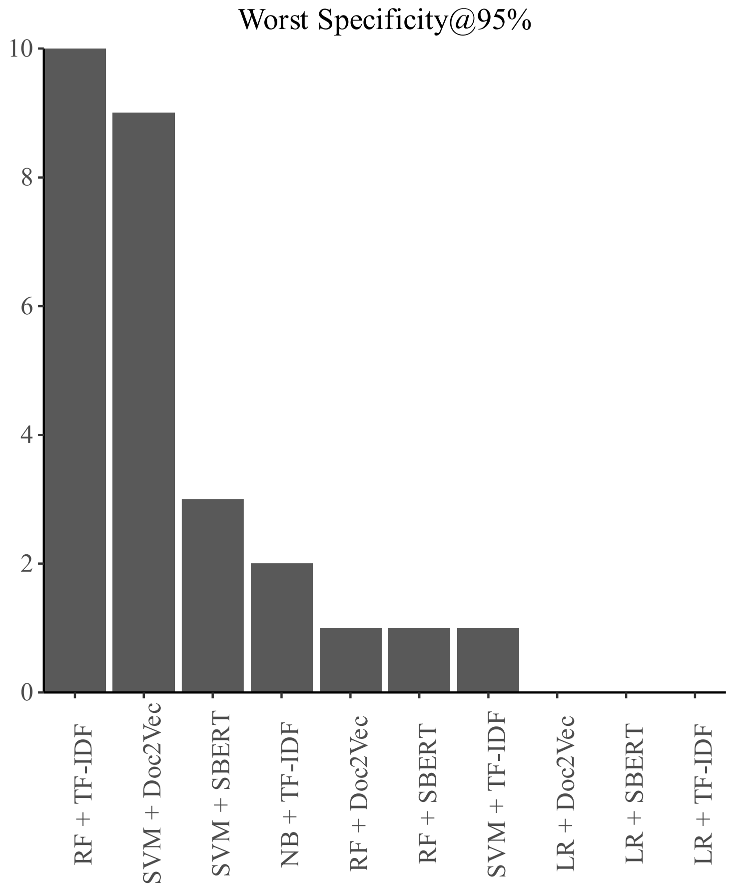
Model Evaluation

Learning Algorithms	Specificity@95%			ETS@95%		
	M	*Mdn*	*SD*	*M*	*Mdn*	*SD*
LR+Doc2Vec	57%	55%	18%	1.62	1.23	1.81
LR+SBERT	65%	71%	18%	1.80	1.29	1.94
LR+TF-IDF	61%	64%	20%	1.75	1.28	1.89
NB+TF-IDF	57%	58%	21%	1.59	1.08	1.77
RF+Doc2Vec	58%	57%	18%	1.65	1.21	1.81
RF+SBERT	63%	68%	19%	1.77	1.39	1.89
RF+TF-IDF	54%	54%	19%	1.54	1.16	1.73
SVM+Doc2Vec	53%	52%	18%	1.57	1.11	1.83
SVM+SBERT	60%	61%	19%	1.71	0.99	1.89
SVM+TF-IDF	57%	58%	20%	1.63	1.22	1.79

Model Evaluation

- The performance of classifiers and feature extraction techniques are similar for educational datasets.
- The RF + SBERT and LR + SBERT models were more frequently identified as the best performing models and produced the highest Specificity@95%.

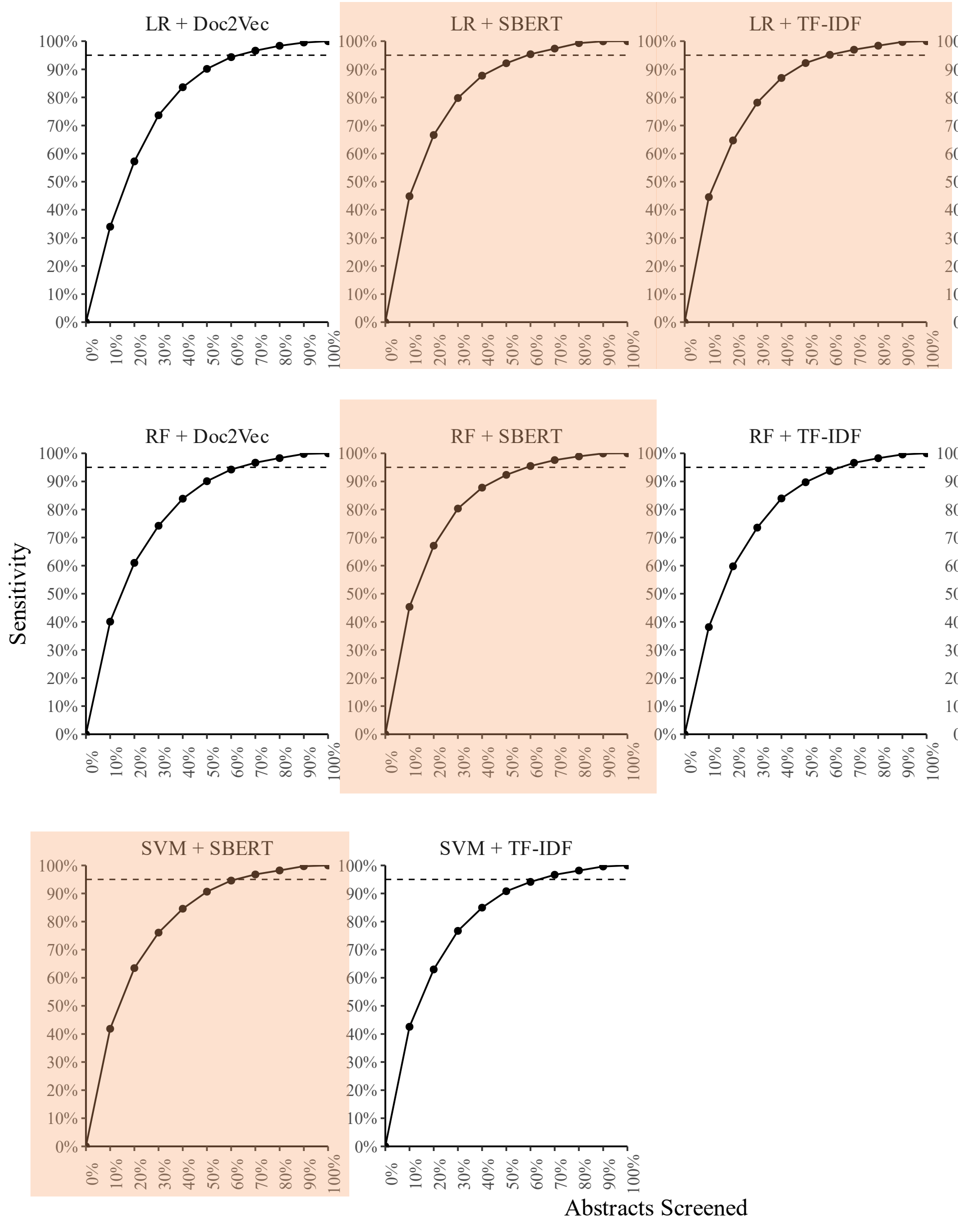




Evaluation of heuristic criteria: Time-based strategy

Percentage reviewed	Sensitivity			Specificity			ETS		
	M	*Mdn*	*SD*	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*
10%	41%	36%	22%	94%	93%	3%	2.57	2.14	2.20
20%	62%	61%	22%	86%	85%	5%	2.28	1.9	1.95
30%	76%	79%	18%	78%	76%	6%	2	1.66	1.71
40%	85%	90%	14%	69%	65%	7%	1.71	1.43	1.46
50%	91%	95%	11%	59%	55%	7%	1.43	1.19	1.22
60%	94%	98%	7%	48%	44%	7%	1.14	0.95	0.98
70%	97%	99%	5%	36%	33%	6%	0.86	0.71	0.73
80%	98%	100%	3%	24%	22%	5%	0.57	0.48	0.49
90%	100%	100%	1%	12%	11%	2%	0.28	0.24	0.24
100%	100%	100%	0%	0%	0%	1%	0	0	0

Evaluation of heuristic criteria: Time-based strategy



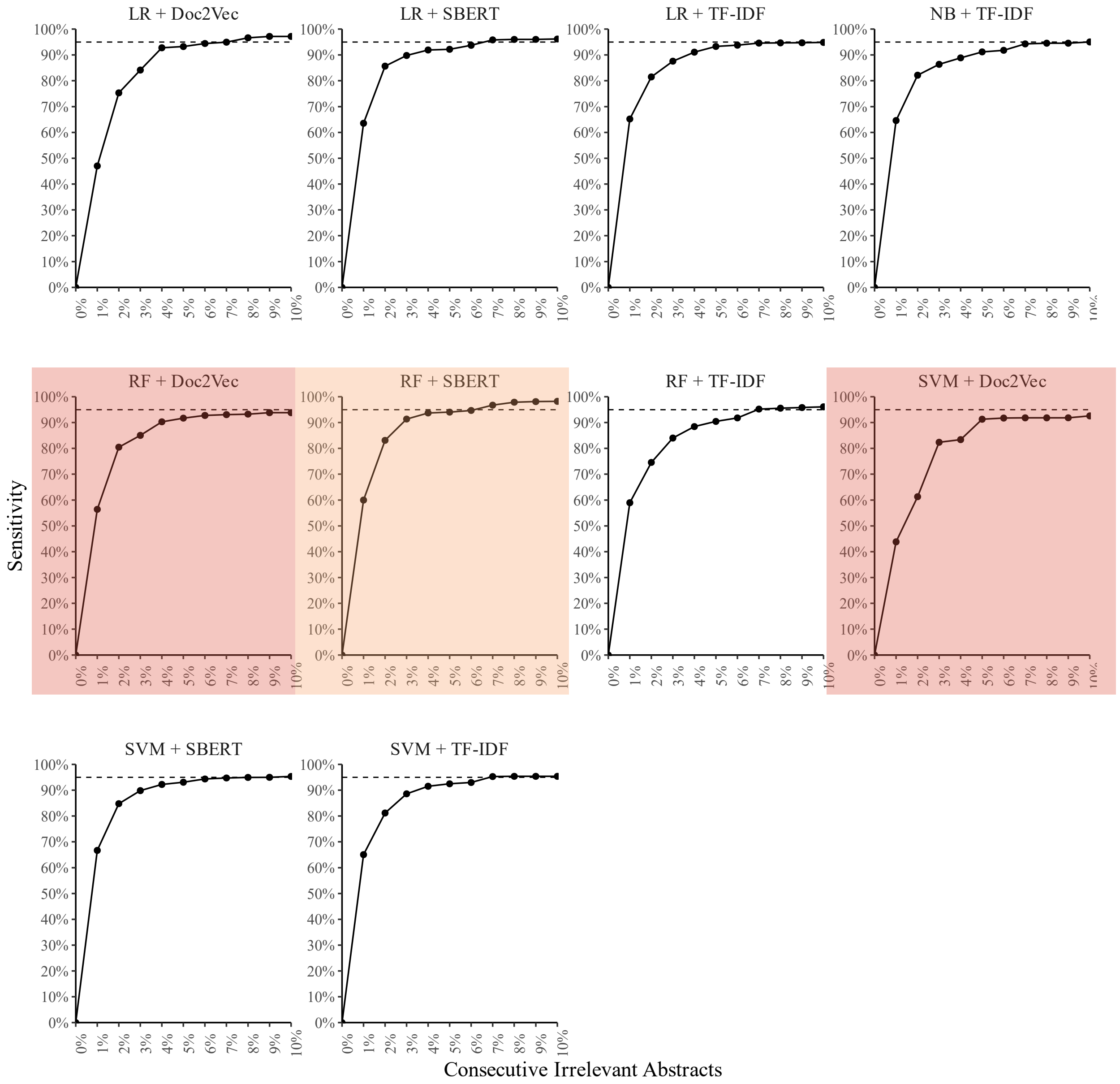
- The time required to find 95% of all relevant records showed minimal differences between the models.
- The performance of the stopping criteria varied greatly between the databases.
  - The database that requires the least amount of screening to find 95% of relevant records is 20%, and the maximum is 90%.
  - The number of records you have to screen to find 95% of all relevant records is positively correlated with the number of relevant records in your database (r = 0.81).

Evaluation of heuristic criteria: Data-based strategy

Percentage irrelevant	Sensitivity			Specificity			ETS		
	M	*Mdn*	*SD*	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*
1%	59%	74%	36%	81%	88%	18%	2	1.32	1.98
2%	79%	93%	29%	65%	71%	25%	1.64	1	1.85
3%	87%	97%	24%	55%	58%	27%	1.43	0.79	1.75
4%	90%	98%	21%	49%	53%	27%	1.28	0.57	1.65
5%	92%	99%	19%	45%	48%	28%	1.16	0.50	1.58
6%	93%	99%	19%	41%	45%	28%	1.09	0.47	1.51
7%	95%	99%	16%	38%	43%	27%	1.04	0.45	1.47
8%	95%	100%	16%	36%	39%	26%	0.97	0.43	1.39
9%	95%	100%	16%	35%	36%	26%	0.93	0.38	1.36
10%	95%	100%	16%	33%	32%	25%	0.89	0.34	1.31

Evaluation of heuristic criteria
Data-based strategy

- There were large variations between different models.
- The performance of the stopping criteria varied greatly between the databases.
  - 1% to 9%.
  - In three of the databases, a cut-down percentage of 10% was insufficient to identify 95% of the records.

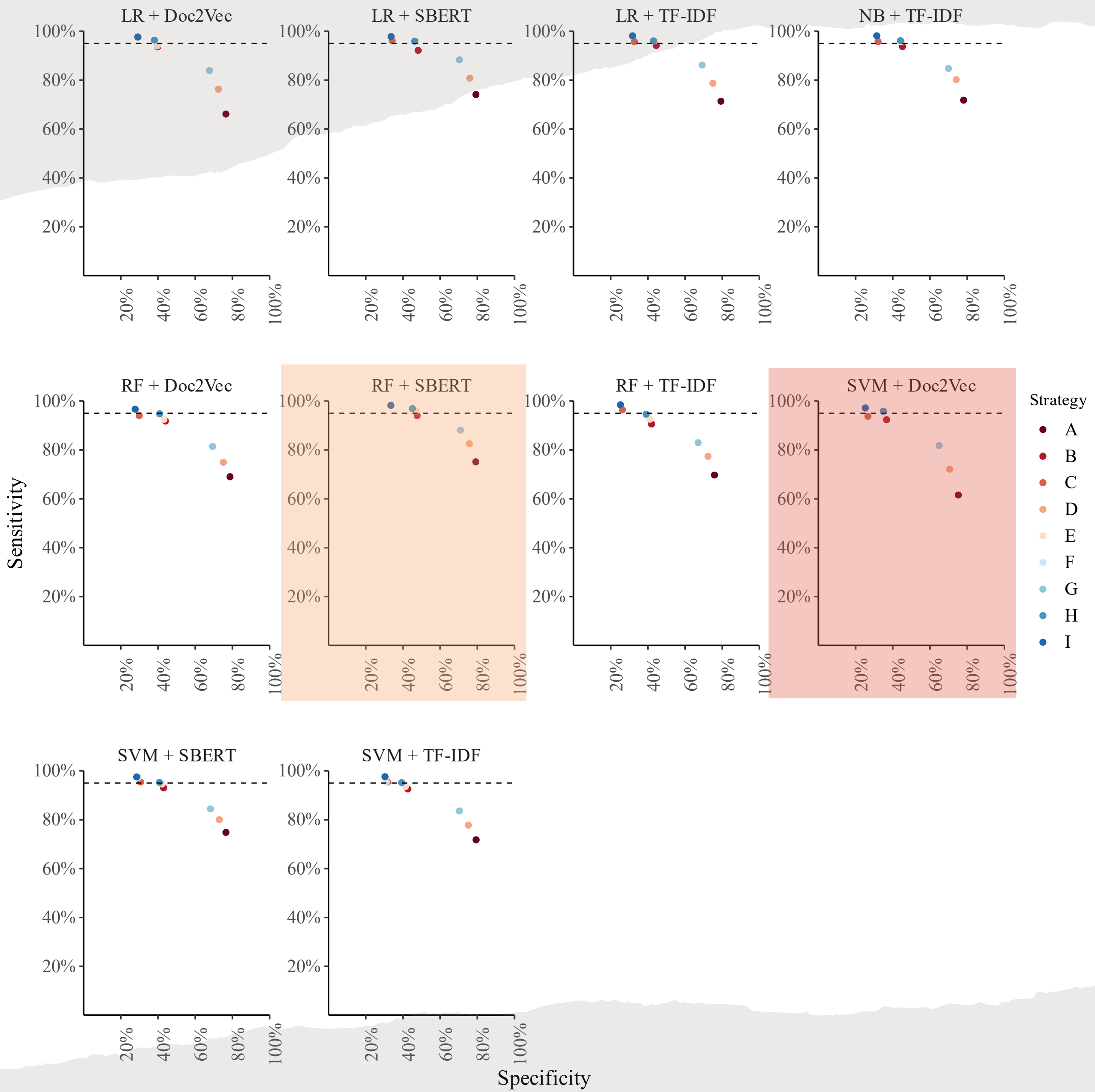


Evaluation of heuristic criteria
Mixed strategy

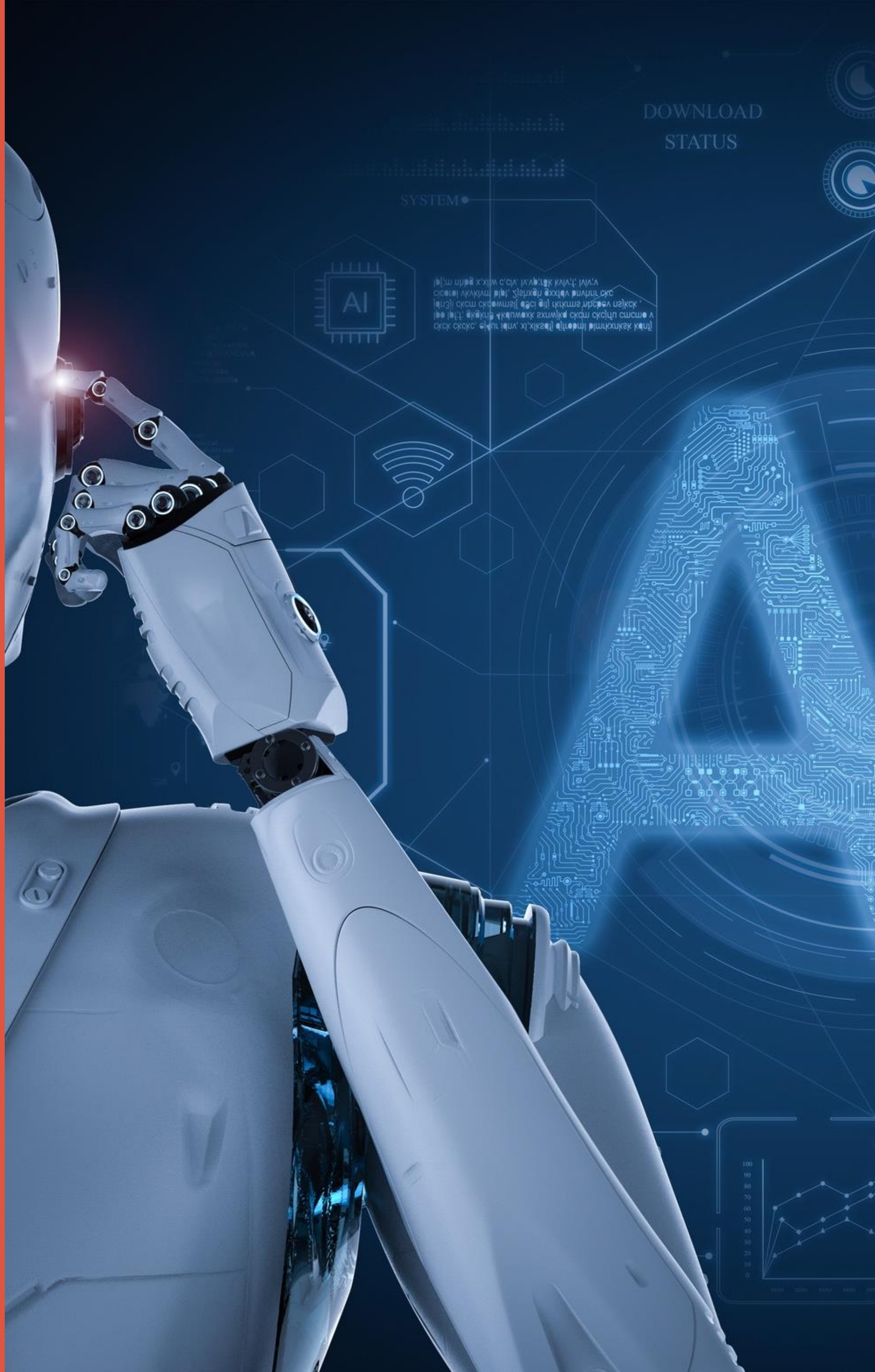
Mixed strategy	Sensitivity			Specificity		
	M	*Mdn*	*SD*	*M*	*Mdn*	*SD*
A	71%	78%	26%	78%	85%	16%
B	93%	99%	19%	43%	48%	29%
C	96%	100%	16%	31%	31%	27%
D	78%	87%	21%	74%	80%	14%
E	95%	99%	13%	42%	48%	28%
F	97%	100%	10%	30%	31%	26%
G	85%	91%	16%	69%	72%	11%
H	96%	99%	11%	41%	48%	26%
I	98%	100%	9%	30%	31%	26%

Evaluation of heuristic criteria
Mixed strategy

- Performance of heuristic strategies varies across models.
- Performance of the models varied depending on the characteristics of the databases and the screening strategies employed.
  - Positive correlations between sensitivity and the number of records.
  - Positive correlation between sensitivity and the number of relevant records.



Conclusion and future research directions



Conclusions

Using active learning for abstract screening allows educational researchers to find relevant records faster than random screening.

The performance of the models varied depending on the characteristics of the databases and the stopping strategies used.

Heuristic-based approaches are heavily dependent on the characteristics of the databases.

It is important to examine how the performance of the models and the stopping criteria changes once more prior information is considered.

Recommendations

1. Data Preparation
- Perform literature search
- Deduplicate search results
- Control format and completeness

2. Random Screening*
- Screen a random subset (e.g. 1%) until at least one relevant abstract is identified
- Store the screening decisions to use the screened abstracts as training set in Step 5

3. Prevalence Estimation
- Estimate prevalence of relevant abstracts based on the random subset (Step 2)
- Alternatively, select an estimation technique for estimating prevalence while screening

4. Model (Learning Algorithm) and Stopping Rule Selection
- Choose model and stopping rules according to the estimated prevalence

Prevalence Ratio of 0.5%

1. LR+doc2vec	Se	Sp	Co
KM	92%	60%	39%
HM	52%	85%	15%
PE	100%	50%	50%

2. LR+SBERT	Se	Sp	Co
KM	85%	79%	21%
HM	63%	88%	12%
PE	100%	58%	42%

Prevalence Ratio of 1%

1. LR+doc2vec	Se	Sp	Co
KM	84%	73%	26%
HM	50%	85%	15%
PE	89%	58%	42%

2. LR+SBERT	Se	Sp	Co
KM	73%	88%	12%
HM	67%	88%	12%
PE	89%	62%	38%

Prevalence Ratio of 5%

1. LR+SBERT	Se	Sp	Co
KM	68%	92%	8%
HM	94%	75%	23%
PE	95%	59%	39%

2. LR+doc2vec	Se	Sp	Co
KM	77%	84%	15%
HM	92%	74%	25%
PE	95%	56%	42%

Prevalence Ratio of 10%

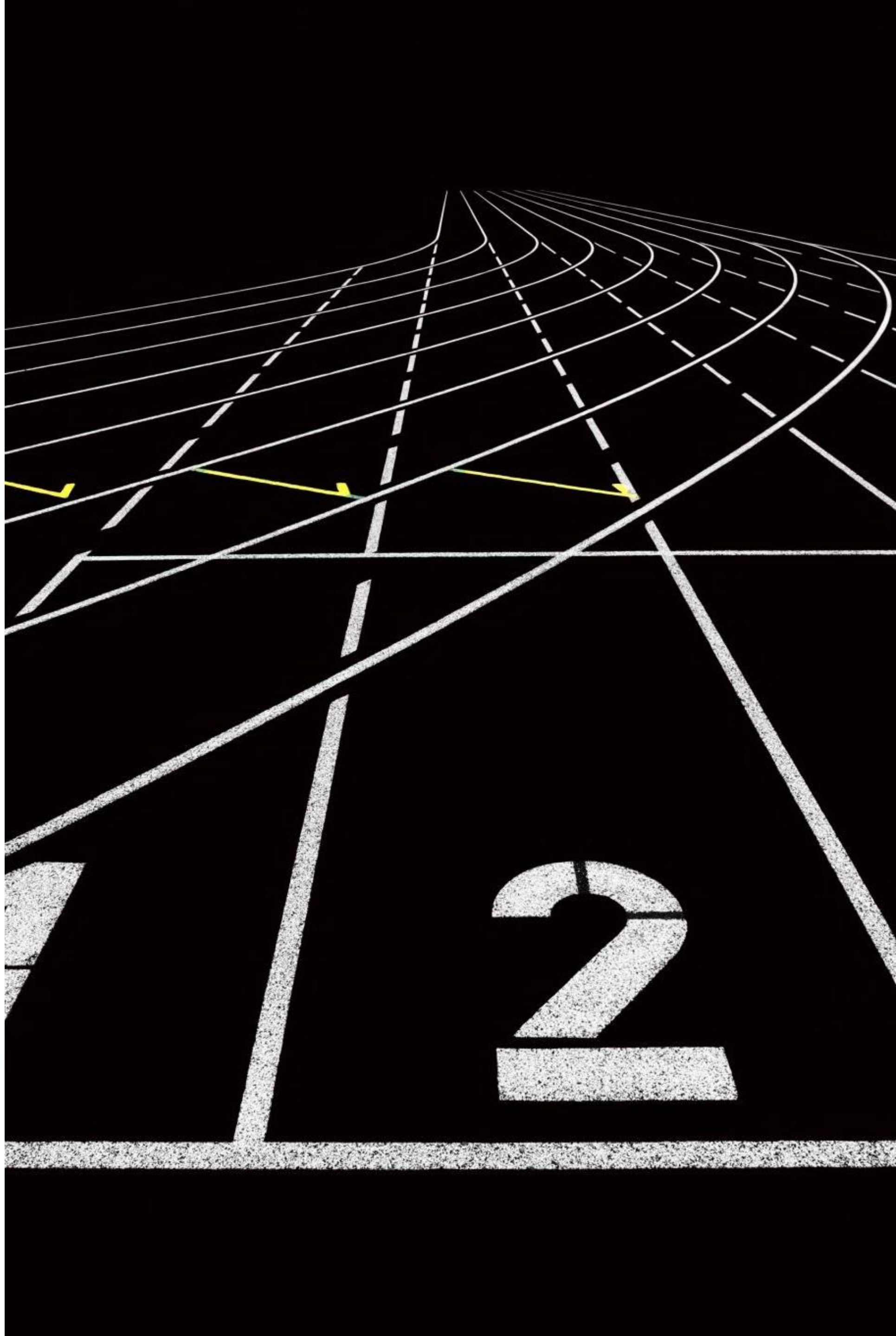
1. LR+SBERT	Se	Sp	Co
KM	73%	90%	9%
HM	96%	70%	28%
PE	95%	59%	38%

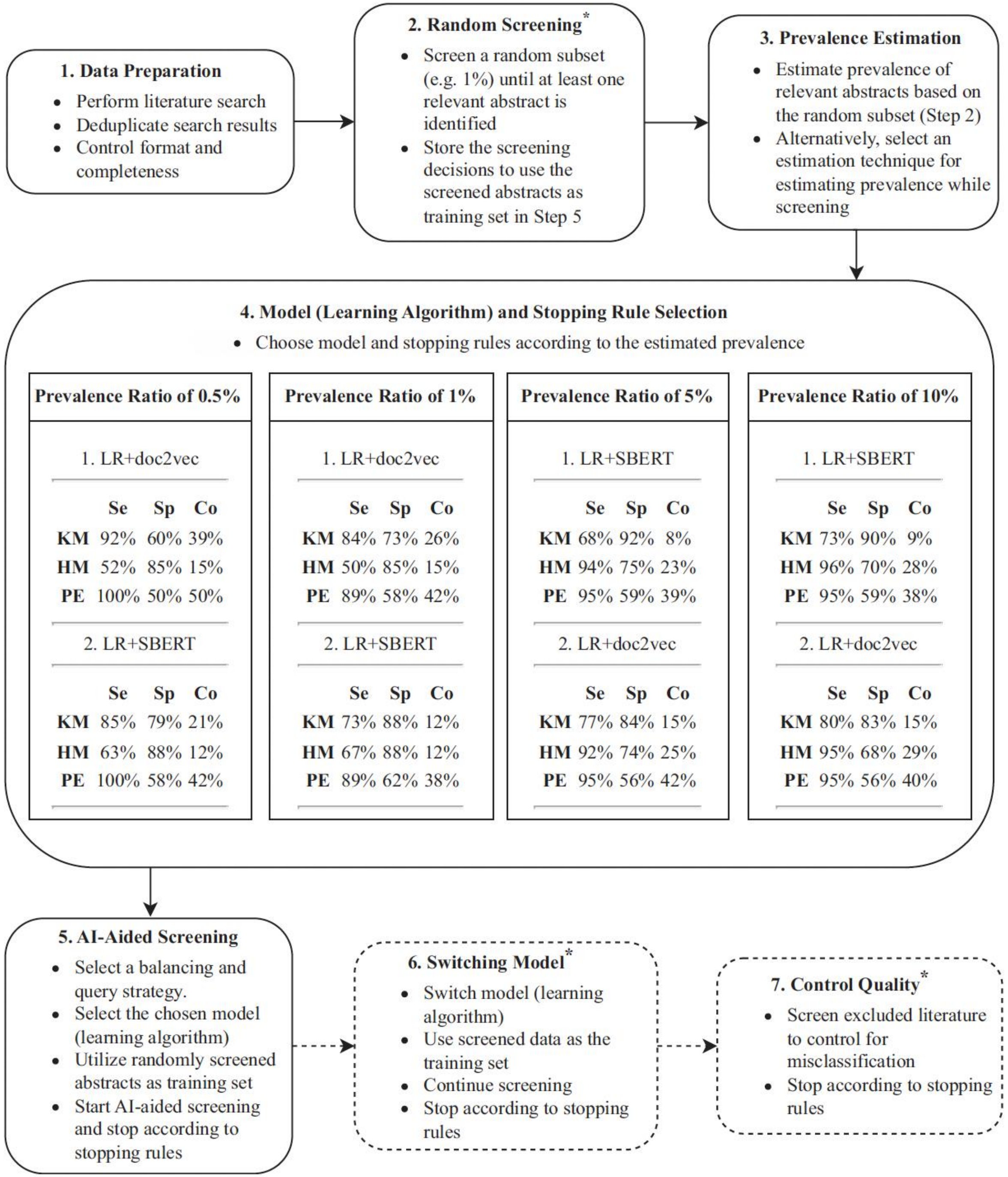
2. LR+doc2vec	Se	Sp	Co
KM	80%	83%	15%
HM	95%	68%	29%
PE	95%	56%	40%

5. AI-Aided Screening
- Select a balancing and query strategy.
- Select the chosen model (learning algorithm)
- Utilize randomly screened abstracts as training set
- Start AI-aided screening and stop according to stopping rules

6. Switching Model*
- Switch model (learning algorithm)
- Use screened data as the training set
- Continue screening
- Stop according to stopping rules

7. Control Quality*
- Screen excluded literature to control for misclassification
- Stop according to stopping rules



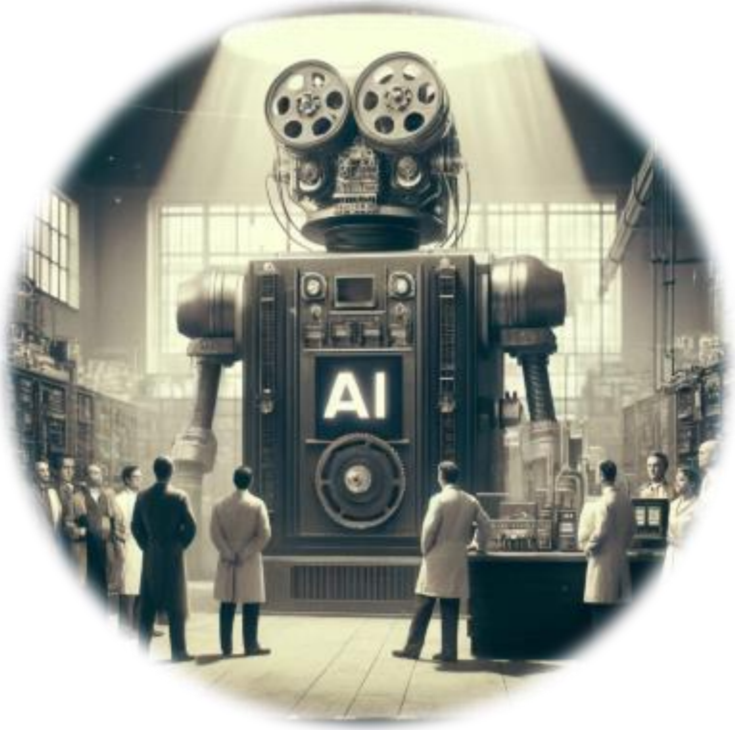

1. Data Preparation
Perform literature search
Deduplicate search results
Control format and completeness
2. Random Screening*
Screen a random subset (e.g. 1%) until at least one relevant abstract is identified
Store the screening decisions to use the screened abstracts as training set in Step 5
3. Prevalence Estimation
Estimate prevalence of relevant abstracts based on the random subset (Step 2)
Alternatively, select an estimation technique for estimating prevalence while screening
4. Model (Learning Algorithm) and Stopping Rule Selection
Choose model and stopping rules according to the estimated prevalence
Prevalence Ratio of 0.5%
1. LR+doc2vec
Se Sp Co
KM 92% 60% 39%
HM 52% 85% 15%
PE 100% 50% 50%
2. LR+SBERT
Se Sp Co
KM 85% 79% 21%
HM 63% 88% 12%
PE 100% 58% 42%
Prevalence Ratio of 1%
1. LR+doc2vec
Se Sp Co
KM 84% 73% 26%
HM 50% 85% 15%
PE 89% 58% 42%
2. LR+SBERT
Se Sp Co
KM 73% 88% 12%
HM 67% 88% 12%
PE 89% 62% 38%
Prevalence Ratio of 5%
1. LR+SBERT
Se Sp Co
KM 68% 92% 8%
HM 94% 75% 23%
PE 95% 59% 39%
2. LR+doc2vec
Se Sp Co
KM 77% 84% 15%
HM 92% 74% 25%
PE 95% 56% 42%
Prevalence Ratio of 10%
1. LR+SBERT
Se Sp Co
KM 73% 90% 9%
HM 96% 70% 28%
PE 95% 59% 38%
2. LR+doc2vec
Se Sp Co
KM 80% 83% 15%
HM 95% 68% 29%
PE 95% 56% 40%
5. AI-Aided Screening
Select a balancing and query strategy.
Select the chosen model (learning algorithm)
Utilize randomly screened abstracts as training set
Start AI-aided screening and stop according to stopping rules
6. Switching Model*
Switch model (learning algorithm)
Use screened data as the training set
Continue screening
Stop according to stopping rules
7. Control Quality*
Screen excluded literature to control for misclassification
Stop according to stopping rules

Challenges and Future Directions

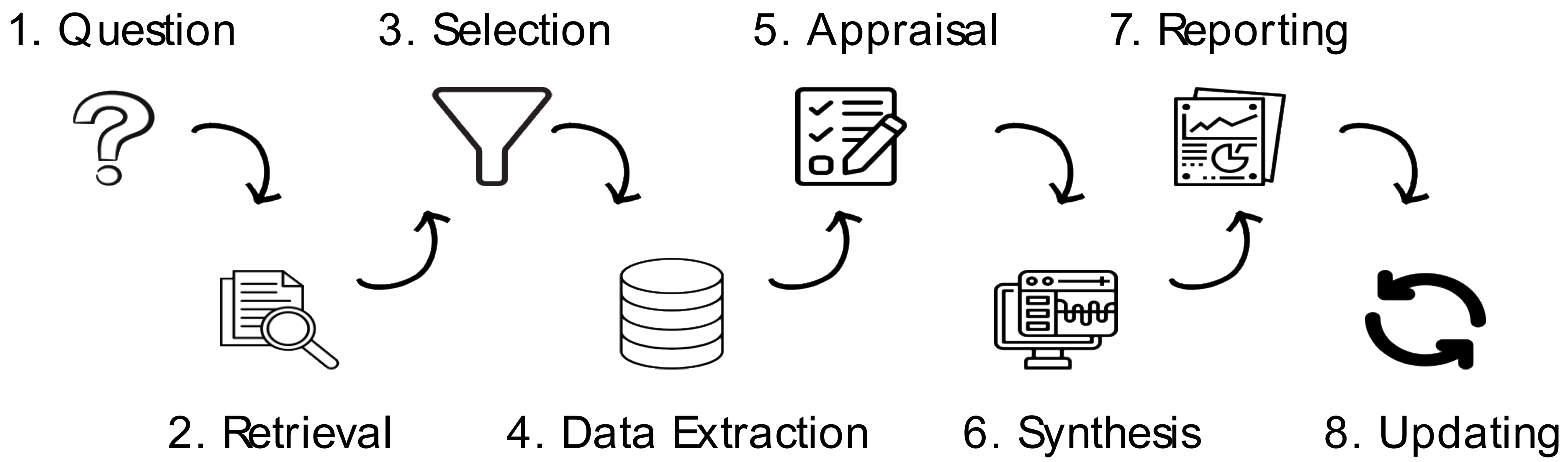
Photo by Jack Millard

AI Systematic Reviews

- Black box algorithms
- Data biases
- Reproducibility of AI workflows



AIRES: Accelerating Research Synthesis in Educational Sciences



(RQ1) How accurate and reliable are AI-based tools for data extraction in research synthesis within the educational field?

(RQ2) What practical steps and ethical considerations should educational researchers follow to ensure both efficiency and scientific rigor when using AI-assisted data extraction?

References

Chai, K. E. K., Lines, R. L. J., Gucciardi, D. F., & Ng, L. (2021). Research Screener: A machine learning tool to semi-automate abstract screening for systematic reviews. *Systematic Reviews*, *10*(1), 93. https://doi.org/10.1186/s13643-021-01635-3

Ferdinands, G., Schram, R. D., Bruin, J. de, Bagheri, A., Oberski, D. L., Tummers, L., & Schoot, R. van de. (2020). *Active learning for screening prioritization in systematic reviews—A simulation study*. OSF Preprints. https://doi.org/10.31219/osf.io/w6qbg

Gates, A., Guitard, S., Pillay, J., Elliott, S. A., Dyson, M. P., Newton, A. S., & Hartling, L. (2019). Performance and usability of machine learning for screening in systematic reviews: A comparative evaluation of three tools. *Systematic Reviews*, *8*(1), 278. https://doi.org/10.1186/s13643-019-1222-2

Gates, A., Johnson, C., & Hartling, L. (2018). Technology-assisted title and abstract screening for systematic reviews: A retrospective evaluation of the Abstrackr machine learning tool. *Systematic Reviews*, *7*(1), 45. https://doi.org/10.1186/s13643-018-0707-8

Olofsson, H., Brolund, A., Hellberg, C., Silverstein, R., Stenström, K., Österberg, M., & Dagerhamn, J. (2017). Can abstract screening workload be reduced using text mining? User experiences of the tool Rayyan. *Research Synthesis Methods*, *8*(3), 275–280. https://doi.org/10.1002/jrsm.1237

Przybyła, P., Brockmeier, A. J., Kontonatsios, G., Le Pogam, M.-A., McNaught, J., von Elm, E., Nolan, K., & Ananiadou, S. (2018). Prioritising references for systematic reviews with RobotAnalyst: A user study. *Research Synthesis Methods*, *9*(3), 470–488. https://doi.org/10.1002/jrsm.1311

Teijema, J., Hofstee, L., Brouwer, M., Bruin, J. de, Ferdinands, G., Boer, J. de, Siso, P. V., Brand, S. van den, Bockting, C., Schoot, R. van de, & Bagheri, A. (2022). *Active learning-based Systematic reviewing using switching classification models: The case of the onset, maintenance, and relapse of depressive disorders*. PsyArXiv. https://doi.org/10.31234/osf.io/t7bpd

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2), Article 2. https://doi.org/10.1038/s42256-020-00287-7

Thank you

Diego G. Campos
diegogc@uio.no
www.diegocampos.co