UNIVERSITEIT UTRECHT

MASTER THESIS

Jump-Starting Evidence Synthesis

Initializing Active Learning Models for Systematic Reviews using LLM-generated Data

Author:

Timo van Ommeren

Supervisors:

Prof. dr. A.G.J. (Rens) van de Schoot, Lauke Stoel

MSc Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Methodology Department, Utrecht University

November 18, 2025
Word count: 2357

1 Introduction

1.1 Background

1.1.1 Active Learning Models for Systematic Reviewing

Researchers and practitioners are continually challenged to base their decisions on the latest scientific evidence. Systematic reviews and meta-analyses were developed to address this need as rigorous methods of summarizing scientific literature (Chalmers, Hedges, and Cooper 2002; Bastian, Glasziou, and Chalmers 2010). However, systematically reviewing the literature can be time-consuming, which limits its practical applicability, especially, in, for example, times of crisis (Tricco et al. 2020; Nussbaumer-Streit et al. 2021).

Fortunately, recent advances in machine learning have produced tools that allow for the systematic screening of scientific literature while greatly reducing the need for manual screening (Van De Schoot et al. 2021). Specifically, active learning models (ALMs) ask users to screen titles and abstracts of papers one by one. Based on the user's decision, the models reassess the probability that the remaining papers are relevant and thus whether to show them to the user. In other words, these models continually reshuffle the papers retrieved from a scientific literature search based on the user's interests. This method reduces the time needed to find as many relevant papers as possible compared to simple index-based screening (Schoot et al. 2025).

1.1.2 The Cold Start Problem

A key challenge to using active learning for systematic reviews is that these models face a "cold start" (Panda and Ray 2022). For an ALM to query a user with a potentially interesting paper, the model must first have knowledge of the user's interests . One way of overcoming a cold start is to initialize, or 'warm up', the ALM using examples of relevant and irrelevant papers (Teijema et al. 2025). If however no examples are available the user may simply start screening 'from the top-down', until a relevant and an irrelevant paper have been found.

1.1.3 The Advent of Large Language Models

With the recent advent of large language models (LLMs), a new possible solution to overcoming the cold start problem has emerged (Bachmann et al. 2025). It may be advantageous to initialize the ALM with synthetic examples of relevant and irrelevant papers, rather than screening from the top-down until actual examples are found. This may be particularly true if the prevalence of relevant papers is rather low. In this case, active learning assisted screening using even a sub-optimal example of a relevant paper may be preferable to random screening for hundreds or thousands of papers. That said, the use of synthetic data may also misdirect the ALM by contaminating the model's training data.

It may therefore be possible to overcome the cold start problem and improve starting performance by using LLMs to generate examples of relevant and irrelevant papers to initialize the ALMs for systematic reviews. This approach could be particularly useful if a user has no relevant examples available, as it avoids top-down screening. However, using synthetic data generated by LLMs seems less likely to improve starting performance if a user has access to actual examples.

1.2 Objectives

This study aims to investigate the effect of using LLM-generated data to initialise active learning models (ALMs) for systematic reviews, compared to random or no initialisation, on starting performance. This will be achieved by simulating the screening of previously published systematic reviews.

2 Methodology

2.1 Simulation set-up

In order to investigate the effect of LLM initialisation, random initialisation and no initialisation on starting performance, the SYNERGY datasets will be screened based on the abstracts and titles of the papers using ASReview's ALM. ASReview is an open-source software package for semi-automated systematic reviewing that implements several ALMs. To simulate the screening process, we will access ASReview via its Python API. For all simulation runs, we will use the recommended U4 configuration in ASReview, which combines a support vector machine classifier with a bi-gram term-frequency inverse document frequency (TF-IDF) feature extractor. The SYNERGY datasets consist of 26 previously screened and labelled sets of papers. Importantly, this gives us access to the ground truth labels of all the papers in these datasets. See Appendix A for a list of all datasets including their topic, total number of records, number of records included and the percentage of relevant records (De Bruin et al. 2023).

2.1.1 How many runs are necessary?

NOTE: Though power estimations don't seem to be standard practice in simulation studies, and a quick reading of the applicable literature (next paragraph) didn't give any indication that many runs would be necessary to estimate starting performance with sufficient precision, it does show metholdogical rigour to do some kind of estimation of Monte Carlo error (Morris, White, and Crowther 2019; Burton et al. 2006). The two options in the literature seem to be either to do a pilot study, or to use an analytical approach. Doing something like this in a systematic manner would be ideal.

Previous simulation studies have found that changing the model parameters in ASReview can significantly effect starting performance, while giving different examples of relevant and irrelevant papers to initialize ASReview has minimal effect on starting performance (Byrne et al. 2024; Teijema et al. 2025). Thus we may conclude that by keeping ASReview's model set to the recommended default U4 configuration over all runs, the effect on starting performance due to random fluctuations should be minimal (J. d. Bruin et al. 2025).

All simulations were done in *Python version 3.10*. For a full list of the packages and their versions, please see the github repository

I am currently aiming for 50 simulation runs per condition per dataset, resulting in a total of 26 (datasets) x 3 (conditions) x 50 (runs) = 3900 simulation runs. (This calculation does not yet consider the variations due to the independent variables in the LLM initialisation condition, so the actual number of runs may be higher).

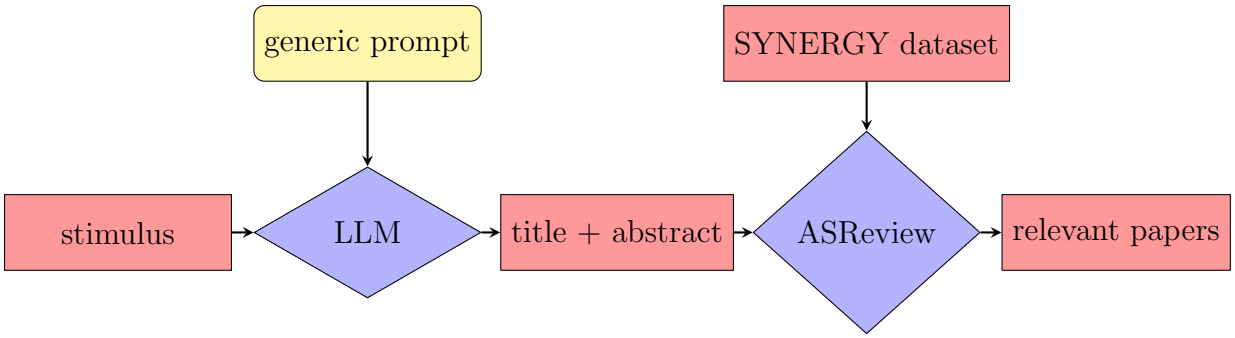

Figure 1: Simulation pipeline (for more detailed schematic of workings within ASReview see (J. d. Bruin et al. 2025))

2.2 Conditions

This section describes how the three study conditions were operationalised. For each run, the ALM parameters are set to the recommended U4 configuration, except for the no initialisation condition switches between two ALMs (see below). The classifier is then seeded with a default seed plus the run number.

See exact code: here

2.2.1 LLM initialization

In the LLM initialisation condition, the ALM's classifier is provided with a set of examples comprising at least one relevant and one irrelevant paper before screening is simulated. These examples are generated based on the inclusion and exclusion criteria of the given systematic review publication. See figure 1 for a schematic of the simulation pipeline.

Between simulation runs the exact number of abstracts generated as well as their specific contents is varied. More specifically, we aimed to investigate the effect of the following variables on starting performance:

1) degree of information provided to the LLM about the systematic review in question **(3 levels)**,
2) number of abstracts generated per simulation run **(4 levels)**,
3) length of abstracts generated per simulation run **(4 levels)**,
4) typicality of abstracts generated per simulation run **(5 levels)**,
5) the use of jargon in generated per simulation run **(5 levels)**.

More specifically, the LLM was provided with either: (a) the title of the published systematic review, (b) the inclusion and exclusion criteria of the systematic review, or (c) the abstract of the published systematic review. Number of abstracts was varied between 1-2-5-10 abstracts per run. The length was varied between 100-200-500-1000 words per run. Whether the abstract should be a typical example of a paper in this review or rather an edge case was varied if 5 or 10 abstracts were generated. In this case the ratio between typical and atypical examples was varied between 00-20-40-60-80-100 % typical. Finally, the use of jargon was varied similarly to typicality by generating abstracts which were simply pure lists of jargon and combining these with regularly written abstract in the ratio 00-20-40-60-80-100 given that at least 5 or 10 abstracts were generated.

To instruct the LLM, a DSPY module was created which takes the variables described above as input, as well as whether the abstract should be relevant or irrelevant (Khattab et al. 2023). Note that to generate multiple abstracts per run, the module was simply looped over and called multiple times. For the exact code, see: here

2.2.2 Random initialization

In the random initialization condition, one relevant and one irrelevant paper were randomly sampled prior to the start of screening. An ALM could therefore be used starting from the very first paper screened.

2.2.3 No initialization

In the no initialisation condition, no papers were provided prior to the start of screening. Therefore, two separate ALMs were used to simulate the screening process. The first ALM randomly sampled papers until at least one relevant and one irrelevant paper were found. These papers were then used to initialise a second ALM, which was used to screen the remaining papers.

2.3 Outcome variable

2.3.1 Starting performance

Starting performance will be assessed based on the number relevant papers in the first 100 papers screened. This figure is derived from research showing that approximately 100 papers can be screened in an hour (Nussbaumer-Streit et al. 2021). For each simulation, we count the number of relevant records found with a *Time to Discovery* below 100 for each simulation (Ferdinands et al. 2023).

2.4 Data Generation and Analysis

2.4.1 Padding

It is important to note that the simulation may stop prematurely if all the relevant records are found before the desired stopping rule is reached (e.g. screening 100 records). In order to accurately emulate a researcher who is unaware that all the relevant records have been found and who therefore continues to screen until the stopping rule is reached, the simulation results were appended with rows containing the label 'zero' (i.e. irrelevant records). The number of these rows equaled the number of records that should have been screened minus the number of records that were screened minus the number of records that were provided prior to screening. This is referred to as padding and ensures that the final simulation results are accurate.

2.4.2 Recall plots

A common way of visualizing the simulation results of retrieval tasks in general, and systematic reviews in particular, are recall curves. These curves shows the number of relevant records retrieved at a given number of papers screened. At the end of each simulation run, a recall plot is created containing three curves: one for each condition. Furthermore, at the end of the entire simulation study, an aggregate recall plot is created

per dataset, containing the average recall curve of each condition, as well as it's standard error between runs.

2.4.3 Exported results

Each simulation run is stored in a separate CSV file. Every row represents a screened paper and contains the following information:
1) The paper's record ID,
2) The assigned label (i.e., relevant or irrelevant),
3) The classifier, querier, balancer and feature extractor used,
4) The size of the training set,
5) A time tag.

Furthermore, the following naming convention is used for the CSV files: condition_run_run_IVs_n_abstracts_length_abstracts_typicality_degree_jargon_llm_temperature.csv. The same naming convention is used for the recall plots of each simulation run and for the generated abstracts in the LLM initialisation condition.

Finally, at the end of each run, the current values of the input parameters, the outcome variables and other relevant metadata are appended to a long format master dataframe for analysis. The columns of this dataframe are as follows:
1) the name of the outcome variable
2) the value of the outcome variable
3) name of the simulated dataset,
4) the condition,
5) the values of the independent variables:
   (a) number of abstracts
   (b) length of the abstracts
   (c) typicality of the abstracts
   (d) degree of jargon in the abstracts
   (e) temperature settings of the llm (i.e., diversity)
6) timestamp
7) the seed value
8) the run number

This yields a data-frame containing one observation for each combination of dataset (n=26), condition (n=3), independent variables and their levels (n=$3\times4\times4\times5\times5 = 1200$), and run (n=1), thus with $26 \times 3 \times 1200 \times 1 = 93600$ rows, and the 12 columns enumerated above.

2.4.4 Analysis

The majority of the variance in the number of relevant records found in the first 100 screened is expected to be explained by differences between datasets. We would therefore like to seperate within-dataset variance from between-dataset variance. Ideal for this purpose are linear mixed-effects models (LMEMs) which allow us to model the effect of condition on starting performance while accounting for between-dataset variance by including dataset as a random effect. To confirm that most of the variance is indeed between datasets, we will visualize the data using boxplots and fit a null model with only dataset as a random effect to calculate the intraclass correlation coefficient (ICC). A

bottom-up modelling approach will then be used for multilevel analysis (for approach see chapter 4 of the book by Hox, Moerbeek, and Van de Schoot 2017) and will be conducted in R using the *lme4* package (Bates et al. 2015).

Question: how to validate exploratory bottom-up approach to modelling. The book mentions either cross-validation or Benferroni correction.

3 Results (note: preliminary!)

3.1 Descriptives

As expected, most of the variance in starting performance seems to be explained by differences between datasets. Figure 2 shows the number of relevant records found within the first 100 screened per dataset. There is considerable variation between datasets, with some datasets yielding less than 10 relevant records found across conditions, while other datasets yield more than 50 relevant records across conditions.

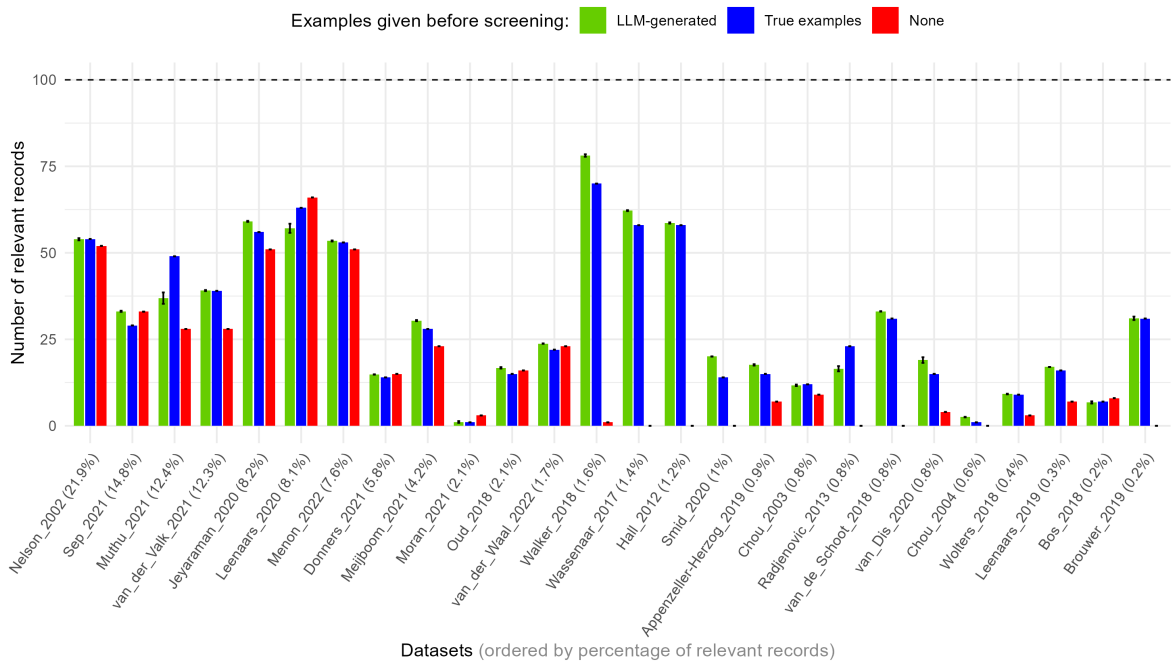

Figure 2: The number of relevant records found per dataset

3.2 Main results

We fit a linear regression to examine whether initialization affected the number of relevant records found within the first 100 screenings. The LLM-initialized model yielded on average B = 30.11, SE = 0.72, t = 41.85. Interestingly, this is approximately equal to number of papers found in the random-initialization condition, at a difference of:, B = -0.08, SE = 1.02, t = -0.08, p = 0.934. In contrast, the difference between the LLM-initialization and the no-initialization conditions was statistically significant: B = -9.88,

$SE = 1.02$, $t = \text{-}9.71$, p ¡ .001. Overall model fit was $R^2 = 0.05$, adjusted $R^2 = 0.04$, and $F(2, 2613) = 62.35$, $p = p$ ¡ .001.

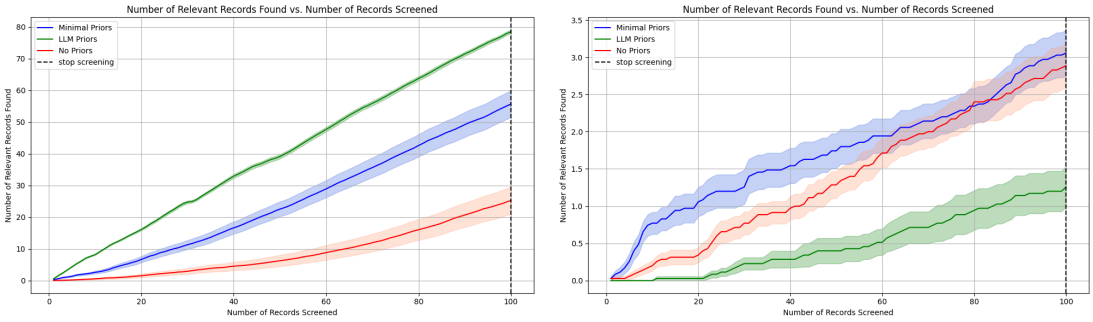

Figure 3: The number of relevant records found per dataset

4 Conclusion

5 Discussion

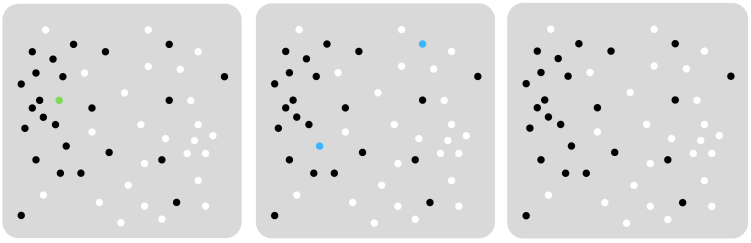

Figure 4: Ideal starting point for systematic reviews using active learning

5.1 Limitations

1. Key limitation simulation study: data leakage (i.e., LLMs have been trained on synergy datasets). The ecological validity of the results are therefore somewhat limited.

   (a) There are two obvious solutions to the problem of data leakage: (1) apply the use of LLm-initialization on a new systematic review, and (2) use an older LLM from hugging face for example.

2. Another possible limitation: no switching of active leaning cycles (could fix contamination of synthetic data issue).

A SYNERGY metadata

References

Bachmann, Fynn et al. (2025). “Adaptive political surveys and GPT-4: Tackling the cold start problem with simulated user interactions”. In: *PLoS One* 20.5, e0322690.

Table 1: Effect of initialization on the number relevant records found in first 100 screened

	Dependent variable:				
	Outcome: number of records found				
	(1)	(2)	(3)	(4)	(5)
conditionminimal		−0.766**	−0.766**	−0.766	−1.562*
		(0.377)	(0.377)	(0.768)	(0.943)
conditionno_priors		−14.420***	−14.420***	−14.420***	−19.421***
		(0.377)	(0.377)	(4.123)	(4.997)
n_abstracts		0.079	0.079	0.079*	0.079*
		(0.188)	(0.188)	(0.040)	(0.040)
length_abstracts		−0.00004	−0.00004	−0.00004	−0.00004
		(0.0005)	(0.0005)	(0.0001)	(0.0001)
llm_temperature		0.121	0.121	0.121	0.121
		(0.451)	(0.451)	(0.097)	(0.097)
Records			0.001**	0.0002	0.0002
			(0.0002)	(0.0002)	(0.0002)
pct			2.294***	2.413***	1.616**
			(0.468)	(0.423)	(0.651)
conditionminimal:pct					0.184
					(0.135)
conditionno_priors:pct					1.159
					(0.715)
Constant	25.820***	30.685***	17.006***	19.132***	22.573***
	(3.464)	(3.509)	(3.808)	(4.442)	(4.851)
Observations	4,212	4,212	4,212	4,212	4,212
Log Likelihood	−16,517.940	−15,750.780	−15,741.930	−9,460.322	−9,458.725
Akaike Inf. Crit.	33,041.880	31,517.560	31,503.870	18,950.640	18,951.450
Bayesian Inf. Crit.	33,060.920	31,568.330	31,567.320	19,045.830	19,059.330
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 2: Datasets overview

Nr	Dataset	Topic(s)	Records	Included	%
1	Appenzeller-Herzog_2019	Medicine	2873	26	0.9
2	Bos_2018	Medicine	4878	10	0.2
3	Brouwer_2019	Psychology, Medicine	38114	62	0.2
4	Chou_2003	Medicine	1908	15	0.8
5	Chou_2004	Medicine	1630	9	0.6
6	Donners_2021	Medicine	258	15	5.8
7	Hall_2012	Computer science	8793	104	1.2
8	Jeyaraman_2020	Medicine	1175	96	8.2
9	Leenaars_2019	Psychology, Chemistry, Medicine	5812	17	0.3
10	Leenaars_2020	Medicine	7216	583	8.1
11	Meijboom_2021	Medicine	882	37	4.2
12	Menon_2022	Medicine	975	74	7.6
13	Moran_2021	Biology, Medicine	5214	111	2.1
14	Muthu_2021	Medicine	2719	336	12.4
15	Nelson_2002	Medicine	366	80	21.9
16	Oud_2018	Psychology, Medicine	952	20	2.1
17	Radjenovic_2013	Computer science	5935	48	0.8
18	Sep_2021	Psychology	271	40	14.8
19	Smid_2020	Computer science, Mathematics	2627	27	1.0
20	van_de_Schoot_2018	Psychology, Medicine	4544	38	0.8
21	van_der_Valk_2021	Medicine, Psychology	725	89	12.3
22	van_der_Waal_2022	Medicine	1970	33	1.7
23	van_Dis_2020	Psychology, Medicine	9128	72	0.8
24	Walker_2018	Biology, Medicine	48375	762	1.6
25	Wassenaar_2017	Medicine, Biology, Chemistry	7668	111	1.4
26	Wolters_2018	Medicine	4280	19	0.4

Bastian, Hilda, Paul Glasziou, and Iain Chalmers (2010). "Seventy-five trials and eleven systematic reviews a day: how will we ever keep up?" In: *PLoS medicine* 7.9, e1000326.

Bates, Douglas et al. (2015). "Fitting linear mixed-effects models using lme4". In: *Journal of statistical software* 67, pp. 1–48.

Bruin, Jonathan de et al. (2025). "ASReview LAB v2: Open-Source Text Screening with Multiple Agents and Oracles". In: *Available at SSRN 5136987*.

Burton, Andrea et al. (2006). "The design of simulation studies in medical statistics". In: *Statistics in medicine* 25.24, pp. 4279–4292.

Byrne, Fionn et al. (2024). "Impact of Active learning model and prior knowledge on discovery time of elusive relevant papers: a simulation study". In: *Systematic Reviews* 13.1, p. 175.

Chalmers, Iain, Larry V Hedges, and Harris Cooper (2002). "A brief history of research synthesis". In: *Evaluation & the health professions* 25.1, pp. 12–37.

De Bruin, Jonathan et al. (2023). "SYNERGY-Open machine learning dataset on study selection in systematic reviews". In: *Version V1*.

Ferdinands, Gerbrich et al. (2023). "Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the Average Time to Discover relevant records". In: *Systematic Reviews* 12.1, p. 100.

Hox, Joop, Mirjam Moerbeek, and Rens Van de Schoot (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Khattab, Omar et al. (2023). "Dspy: Compiling declarative language model calls into self-improving pipelines". In: *arXiv preprint arXiv:2310.03714*.

Morris, Tim P, Ian R White, and Michael J Crowther (2019). "Using simulation studies to evaluate statistical methods". In: *Statistics in medicine* 38.11, pp. 2074–2102.

Nussbaumer-Streit, Barbara et al. (2021). "Resource use during systematic review production varies widely: a scoping review". In: *Journal of clinical epidemiology* 139, pp. 287–296.

Panda, Deepak Kumar and Sanjog Ray (2022). "Approaches and algorithms to mitigate cold start problems in recommender systems: a systematic literature review". In: *Journal of Intelligent Information Systems* 59.2, pp. 341–366.

Schoot, Rens van de et al. (2025). "The Hunt for the Last Relevant Paper: Blending the best of humans and AI". In: *European Journal of Psychotraumatology* 16.1, p. 2546214.

Teijema, Jelle Jasper et al. (2025). "Large-scale simulation study of active learning models for systematic reviews". In: *International Journal of Data Science and Analytics*, pp. 1–22.

Tricco, Andrea C et al. (2020). "Rapid review methods more challenging during COVID-19: commentary with a focus on 8 knowledge synthesis steps". In: *Journal of clinical epidemiology* 126, pp. 177–183.

Van De Schoot, Rens et al. (2021). "An open source machine learning framework for efficient and transparent systematic reviews". In: *Nature machine intelligence* 3.2, pp. 125–133.