



Efficacy of Artificial Intelligence in Reducing Number of Articles Screened in Systematic Reviews: An ICAROS II post-hoc analysis.

Periklis Giannakis MD¹, Stavros G. Memtsoudis MD MBA PhD^{1,3}, Crispiana Cozowicz MD², Jiabin Liu MD PhD FASA^{1,3}, Lisa Reisinger MD^{1,2}, Alex Illescas MPH¹, Ellen Soffin MD PhD^{1,3}, Mary Hargett¹, Bridget Jivanelli MLIS¹, Juliet Rowe MPH¹, Jashvant Poeran MD PhD¹

¹Department of Anesthesiology, Critical Care & Pain Management, Hospital for Special Surgery, New York, NY, USA
²Department of Anesthesiology, Perioperative Medicine and Intensive Care Medicine, Paracelsus Medical University, Salzburg, Austria
³Weill Cornell Medicine, New York, NY, USA



INTRODUCTION

Significance: **SRs**→ qualitative/quantitative assessment and synthesis of existing evidence→ **imperative for medical literature**

The problem: Increasingly time-consuming given the ever-increasing amount of literature combined with adoption of stricter methodological requirements

The solution (?): AI offers a promising, Cochrane-approved solution to make this process more efficient

In this post-hoc analysis: Data from a recently finalized SR by our group (ICAROS II [PROSPERO: CRD42024475582], ASRA 50 #6865), we tested an AI approach and determined its efficacy

METHODS

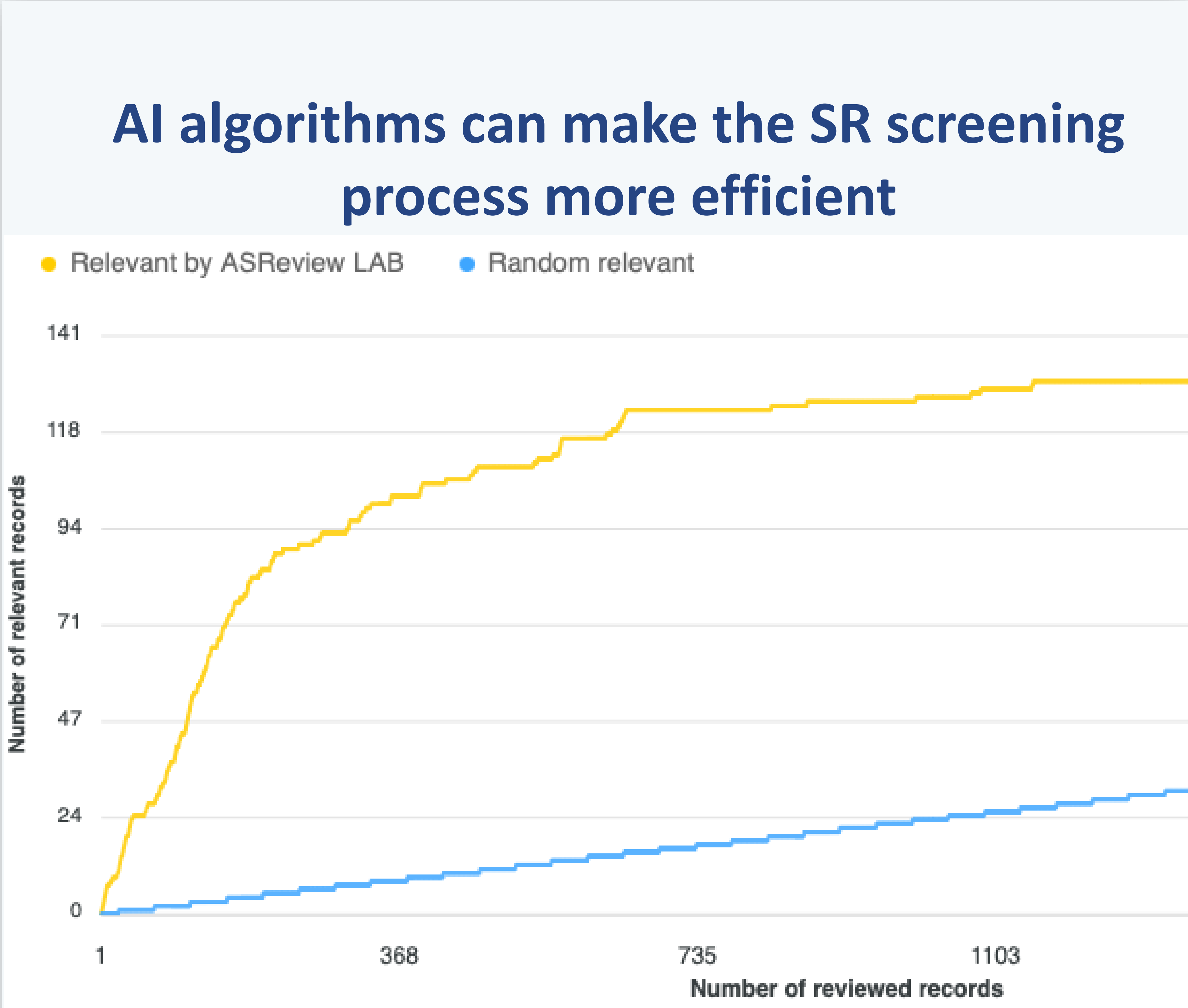
Comparison: Manual vs AI-assisted approach (AS Review Lab)

Data collection: A deduplicated .csv file of all studies included in titles and abstract screening was uploaded to AS Review Lab. A .csv file with all the included studies for data extraction was also downloaded to be used as a reference

Active learning model set-up: AS Review Lab active learning model consists of; a feature extraction technique, a classifier, a query strategy and a balance strategy
We opted for the default setup (TD-IDF,NB, Maximum, Dynamic resampling) as suggested by the creators

Prior knowledge: 2 relevant and 2 irrelevant articles, handpicked by the screener/reviewer (PG)

Outcomes: Primary outcome: total labeled records (relevant/irrelevant), until all articles included in data extraction were labeled as relevant
Secondary outcome: maximum number of consecutive irrelevant records since last relevant



However, algorithms are influenced by numerous factors that may mitigate this initial efficiency

RESULTS

PRISMA relevant highlights: Studies screened (n=6,247), studies included in review (n=133)

Primary outcome: We screened n=1,363 studies (21.8%) to retrieve all studies included in the review

Secondary outcome: The maximum number of consecutive irrelevant records since last relevant was n=195, and happened between the 131st and 132nd relevant studies

DISCUSSION

The utilized AI tool considerably reduced the number of articles in the screening process: However, 2 studies may have been missed if we had opted to use predetermined stopping criteria reported, i.e. to stop after a certain amount of consecutive irrelevant studies

We only needed to go through 20% of the articles, but there are several study-related variables that heavily influence algorithm performance: Factors such as the research question, inclusion/exclusion criteria, characteristics of articles retrieved (number, relevance, abstract available in .csv file, etc.) massively influence the algorithm as TD-IDF/NB can overemphasize rare but irrelevant terms, leading to early “false positives” and late “false negatives”

Other limitations: TD-IDF/NB’s inability to predict dataset balance (relevant/irrelevant studies ratio) in its decision-making algorithm, and insufficient accounting for semantics, interchangeable terms and context

In conclusion: AI algorithms can make the SR screening process more efficient, however, algorithms are influenced by numerous factors that may mitigate this initial efficiency

More Info →

