第9回研究倫理教育研修会
2025年5月1日
日本医師会館

AIを用いたガイドライン作成

大網 毅彦
千葉大学大学院医学研究院　救急集中治療医学

第9回研究倫理教育研修会

利益相反（COI）開示

筆頭演者氏名：大網　毅彦

演題発表に関連し，開示すべきCOI関係にある企業などはない。

本日の内容

- 自己紹介
- 日本版敗血症診療ガイドライン
- ガイドライン作成における作業負担の大きさ
- 文献スクリーニング自動化ソフト
- 生成AIを用いた文献スクリーニング
- AIを用いたガイドライン作成の展望

自己紹介

大網 毅彦
救急科専門医・指導医
集中治療専門医

略歴

2006年 東京慈恵会医科大学医学部卒業
2008年 千葉大学 救急集中治療医学
2016年 千葉大学大学院卒業（医学博士）
2018年 米国Emory Universityへ留学
2020年 千葉大学 救急集中治療医学 助教
2023年 千葉大学 救急集中治療医学 講師

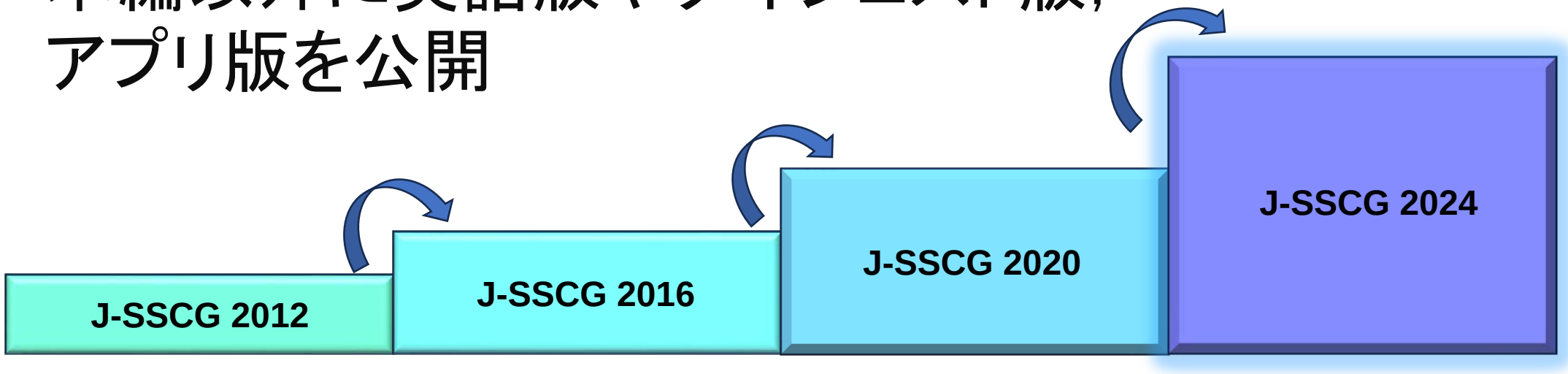
ガイドラインへの関わり

日本版敗血症診療ガイドライン委員長補佐

日本版敗血症診療ガイドライン

日本版敗血症診療ガイドライン（J-SSCG）の概要

- 2012年から4年ごとに改訂
- 2024年12月に正式公開
- 本編以外に英語版やダイジェスト版，アプリ版を公開



J-SSCG2024のポイント

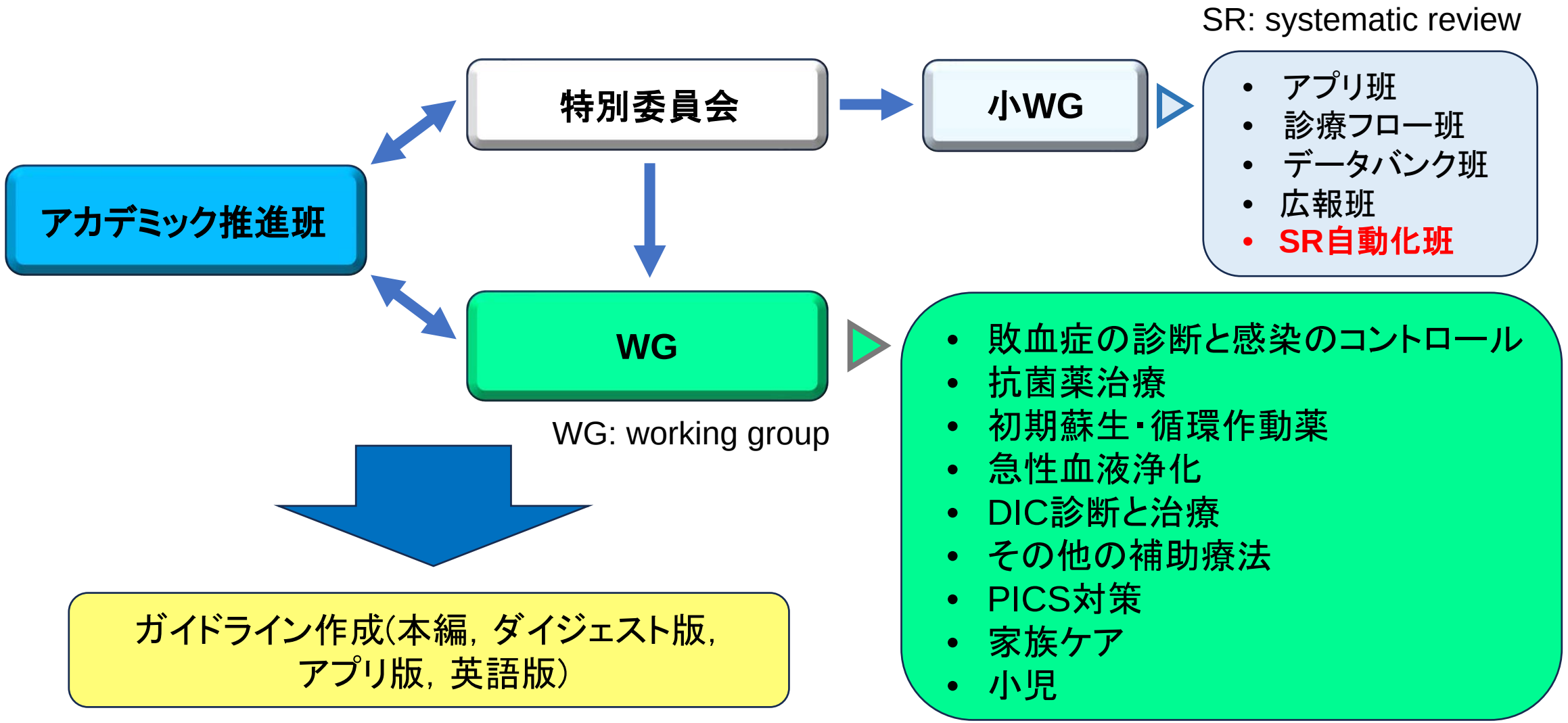
1. より使いやすいガイドラインの作成

2. 効率化・負担軽減，最先端の作成手法の導入

3. ガイドライン作業におけるエビデンスの創出

4. ガイドライン作業を通じた人材発掘と育成

J-SSCG2024の組織構成



ガイドライン作成における作業負担の大きさ

診療ガイドライン作成における systematic review

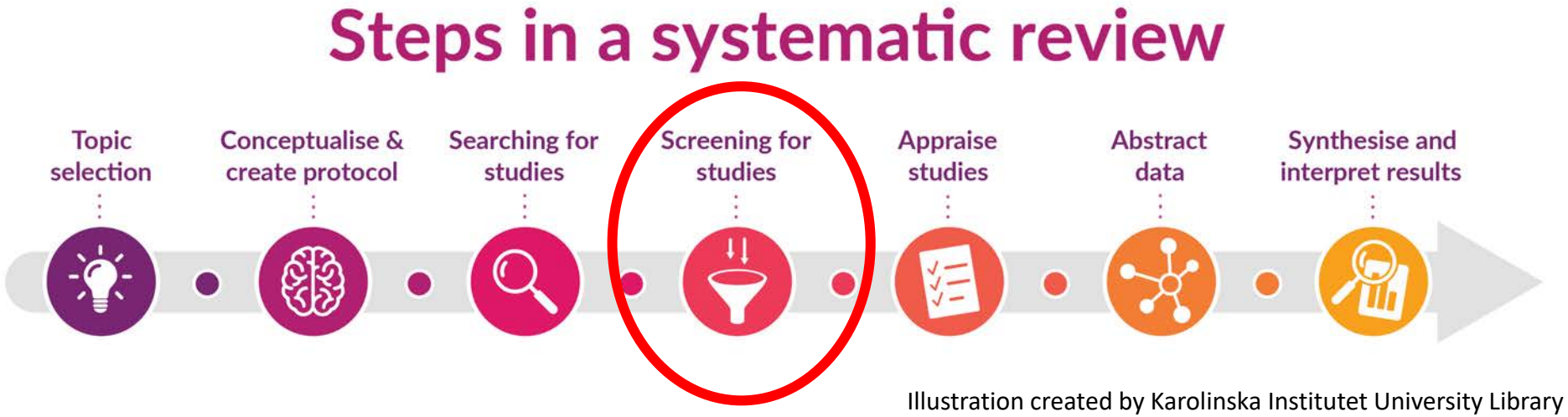


Illustration created by Karolinska Institutet University Library

診療ガイドラインでは多くのclinical question (CQ)に対する文献スクリーニングを一定期間で進めていく必要があり，ガイドライン作成メンバーの作業は膨大かつ負担が大きい。

SRの作業負担とエラー出現に関する過去の報告

✓システマティックレビューの実施には平均67週間を要し，かつ多くの人的資源が必要であった。

Borah R, et al. BMJ Open. 2017;7(2):e012545.

✓システマティックレビューにおける文献スクリーニングにおいて，約10.8%に何らかのエラーを認めた。

Wang Z, et al. PLoS One. 2020;15(1):e0227742.

持続可能なガイドライン作成を行なっていくためには，作業の効率化や負担軽減を行なっていく必要がある。

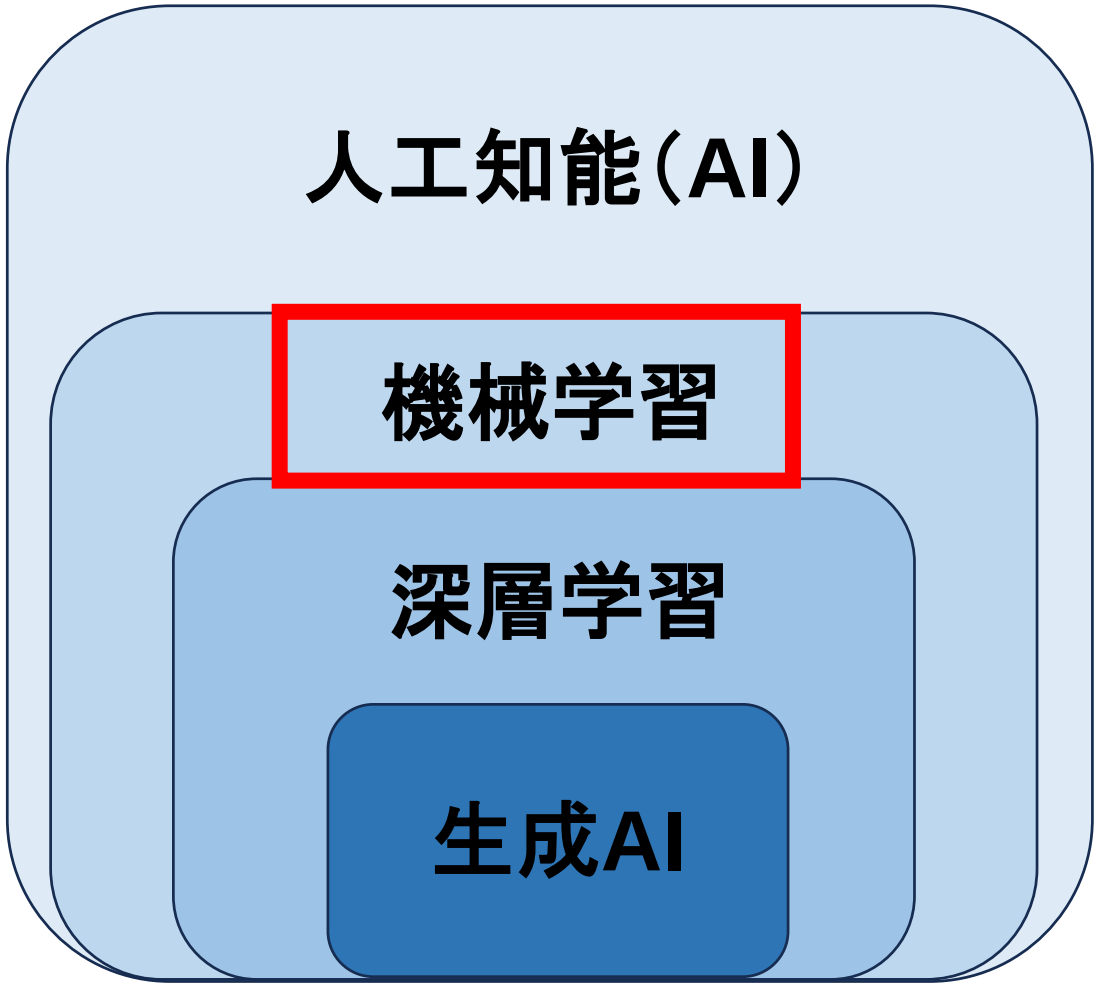
SR自動化班の目標

SRの自動化を進めることで診療ガイドライン作成のための作業量の削減を目指す。

人工知能（AI）がガイドライン作業の一部を代替できる可能性あり

文献スクリーニング自動化ソフト

AIの分類



人間のように学習・判断・推論する能力を持つコンピュータシステムの総称

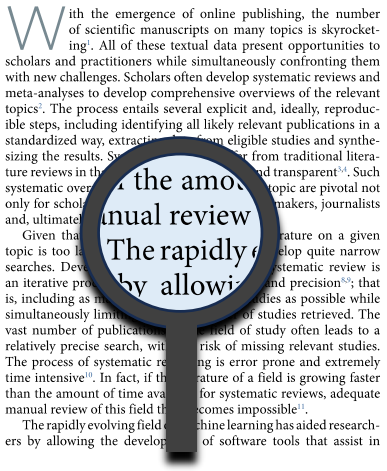
データからパターンを学習して予測や分類を行う技術

人間の脳にヒントを得たニューラルネットワークを使って複雑な特徴を自動で学習する方法

文章や画像，音声などの新しいデータを人間のように生成する技術

AIを用いた文献スクリーニングの自動化

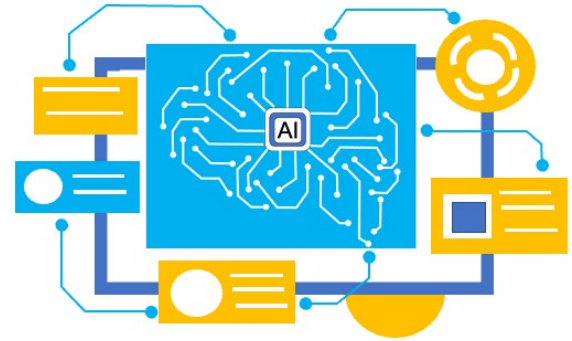
文献


With the emergence of online publishing, the number of scientific manuscripts on many topics is skyrocketing[1]. All of these textual data present opportunities to scholars and practitioners while simultaneously confronting them with new challenges. Scholars often develop systematic reviews and meta-analyses to develop comprehensive overviews of the relevant topics[2]. The process entails several explicit and, ideally, reproducible steps, including identifying all likely relevant publications in a standardized way, extracting [illegible] from eligible studies and synthesizing the results. Sy[illegible]r from traditional literature reviews in th[illegible]d transparent[3,4]. Such systematic over[illegible]topic are pivotal not only for schol[illegible]makers, journalists and, ultimatel[illegible]

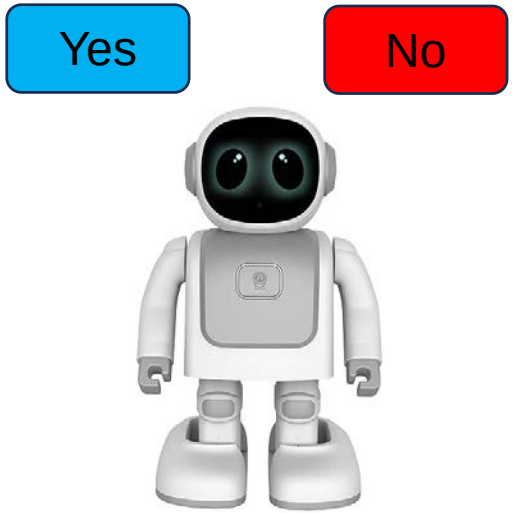
Given tha[illegible]rature on a given topic is too l[illegible]elop quite narrow searches. Deve[illegible]ystematic review is an iterative pro[illegible]and precision[4,5]; that is, including as m[illegible]dies as possible while simultaneously limi[illegible]of studies retrieved. The vast number of publication[illegible]ield of study often leads to a relatively precise search, wit[illegible]risk of missing relevant studies. The process of systematic re[illegible]ng is error prone and extremely time intensive[10]. In fact, if th[illegible]rature of a field is growing faster than the amount of time ava[illegible]for systematic reviews, adequate manual review of this field th[illegible]comes impossible[11].

The rapidly evolving field [illegible]chine learning has aided researchers by allowing the develop[illegible]of software tools that assist in

the amo[illegible]
nual review
The rapidly
by allowi[illegible]



PICOに合致した文献か？



文献スクリーニング自動化ソフトは学習データをもとに CQのPICOに合致した文献を自動的に探し出す。

PICO: patient/population, intervention, comparison, outcomes

文献スクリーニングの自動化により予想されるメリットとデメリット

メリット

スクリーニング時間の短縮による作業負担の軽減

デメリット

文献スクリーニングの精度の担保

自動化文献スクリーニングソフト

診療ガイドライン作成のための半自動化文献スクリーニングソフトの精度と作業負荷軽減の検証：前向き観察研究

目的

日本版敗血症診療ガイドラインの作成において，半自動化文献スクリーニングソフトの精度と作業負荷軽減の評価を行うこと。

方法

1. 診療ガイドライン作成に適した半自動化スクリーニングソフトを選定する。

2. 2人の独立したreviewerが自動化スクリーニングソフトを用いてJ-SSCG2024のCQに対する1次スクリーニングを行い，同時期に従来の方法で行われたスクリーニング結果と比較する。

Primary outcome: 自動化文献スクリーニングの精度

Secondary outcome: スクリーニングに要した時間

本研究に用いたCQ一覧

- 敗血症の初期蘇生における低灌流の指標は?
- 循環動態が安定した敗血症に対して制限的輸液管理を行うか？
- 敗血症に対する初期輸液にどの輸液製剤を用いるか？
- 敗血症に対する初期蘇生において，平均動脈圧の目標値をいくらとするか？
- 重度の代謝性アシドーシス（pH≦7.2）を伴う敗血症に対して，重炭酸ナトリウムの静脈投与を行うか?

半自動化文献スクリーニングソフトの選定

39 Software assessed for eligibility

35 Did not meet inclusion criteria

Inclusion criteria

- Open source
- Machine learning
- Active learning

3 Excluded

1 Slow processing speed

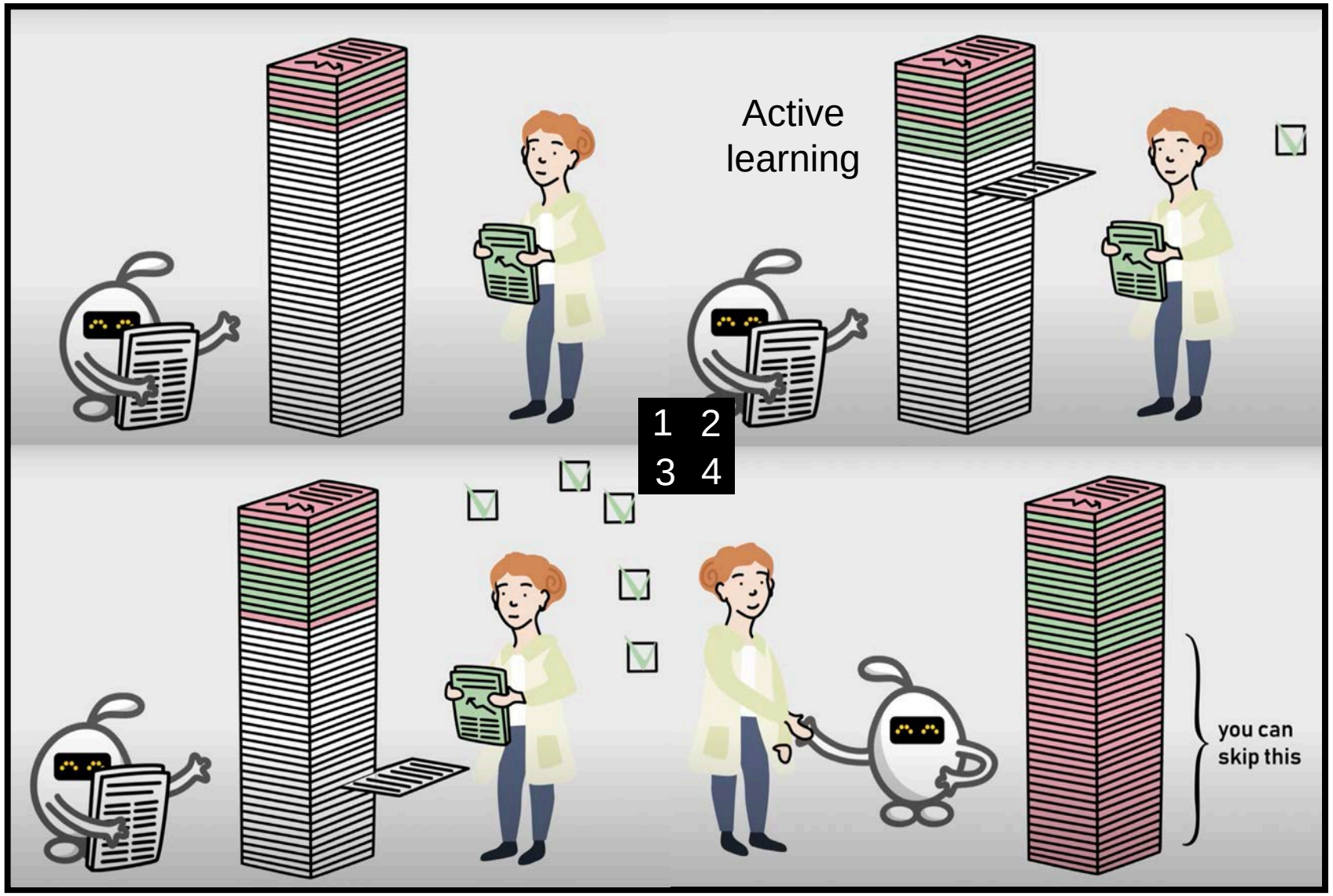
1 Incompatibility with other software

1 Difficulty in application settings

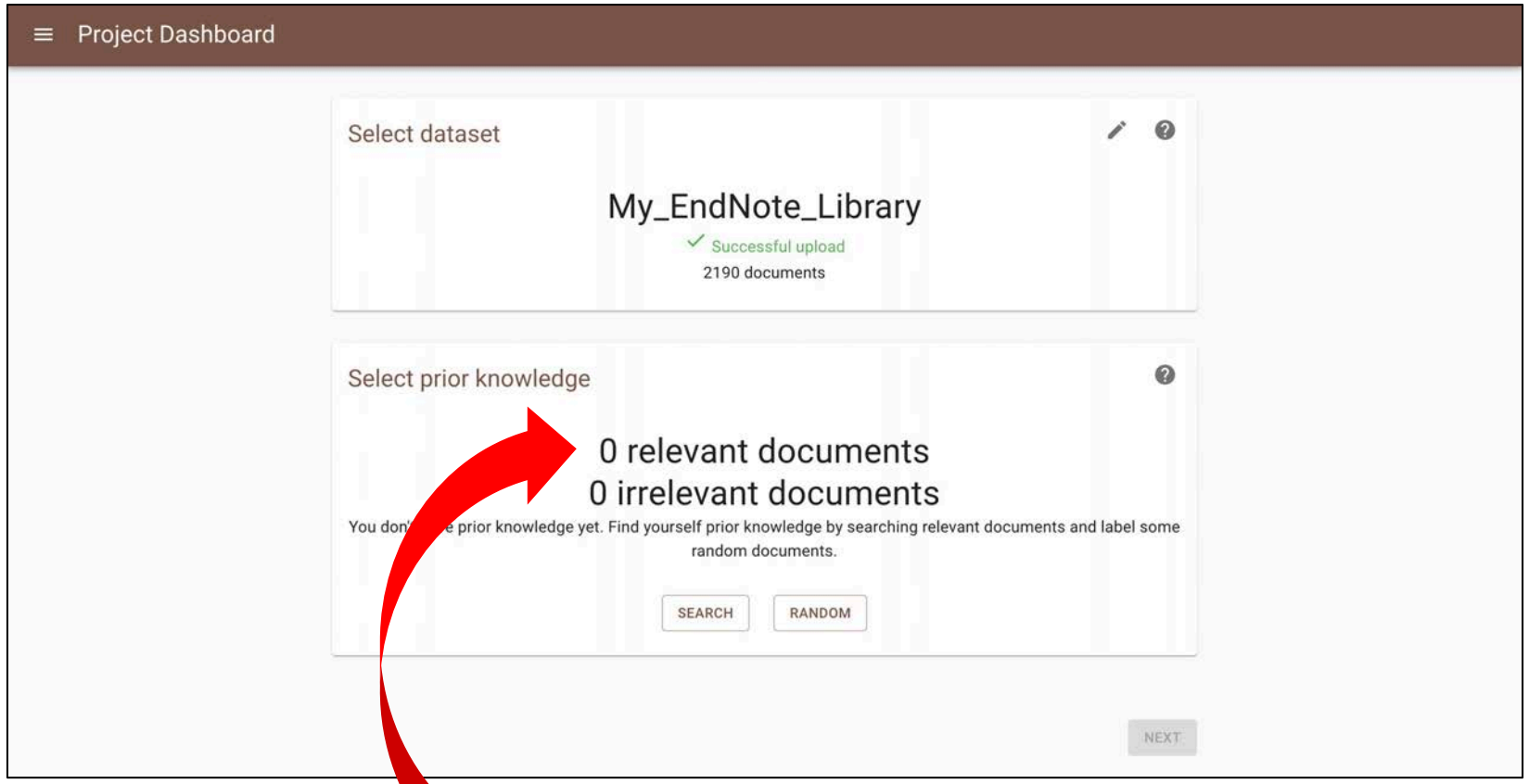
1 Included in the study

ASReview

ASReview

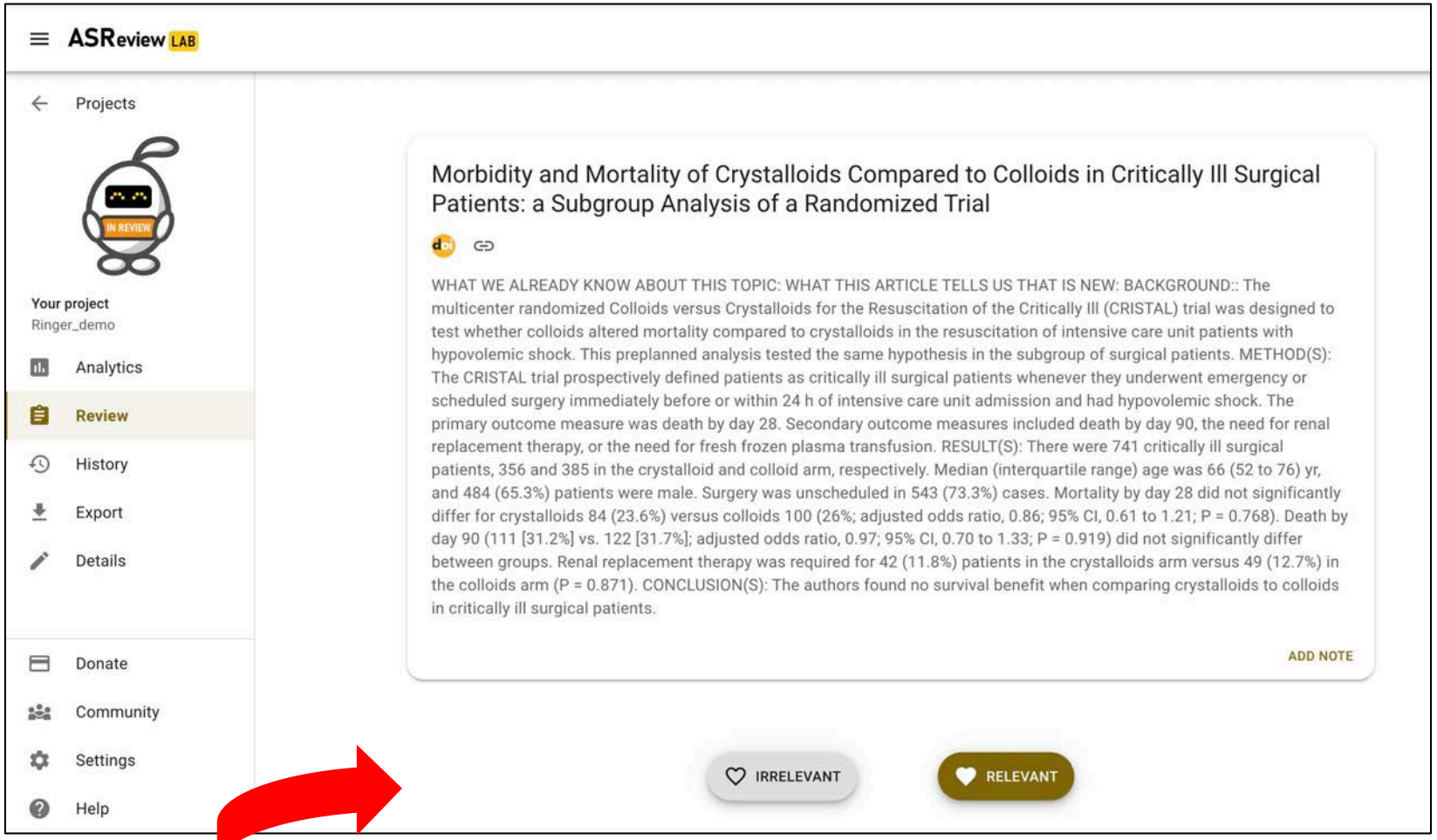


ASReviewの操作画面



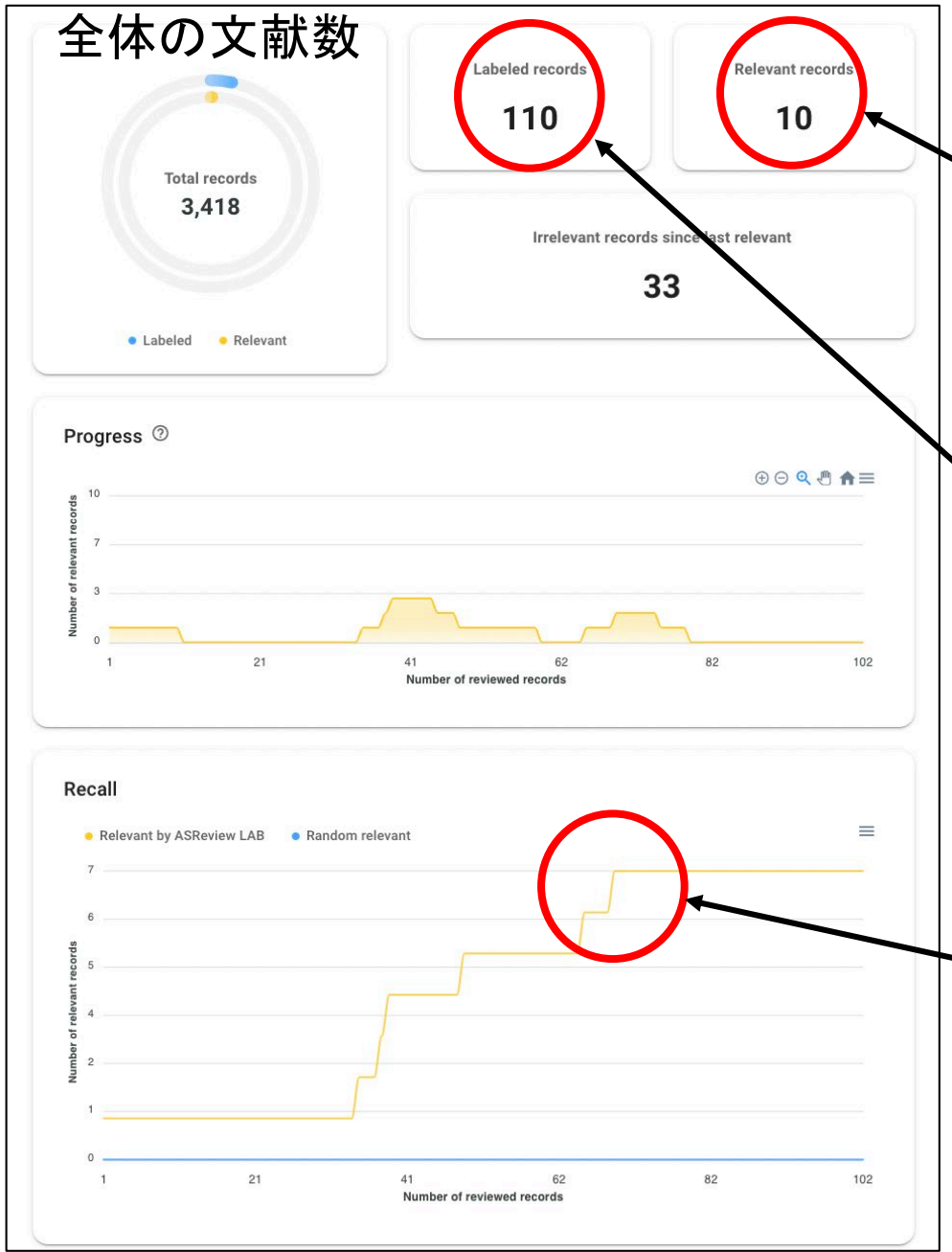
学習データとしてキー論文を入力する

ASReviewの操作画面



追加の学習データとしてスクリーニングを行う。

ASReviewを用いた文献スクリーニングの実際

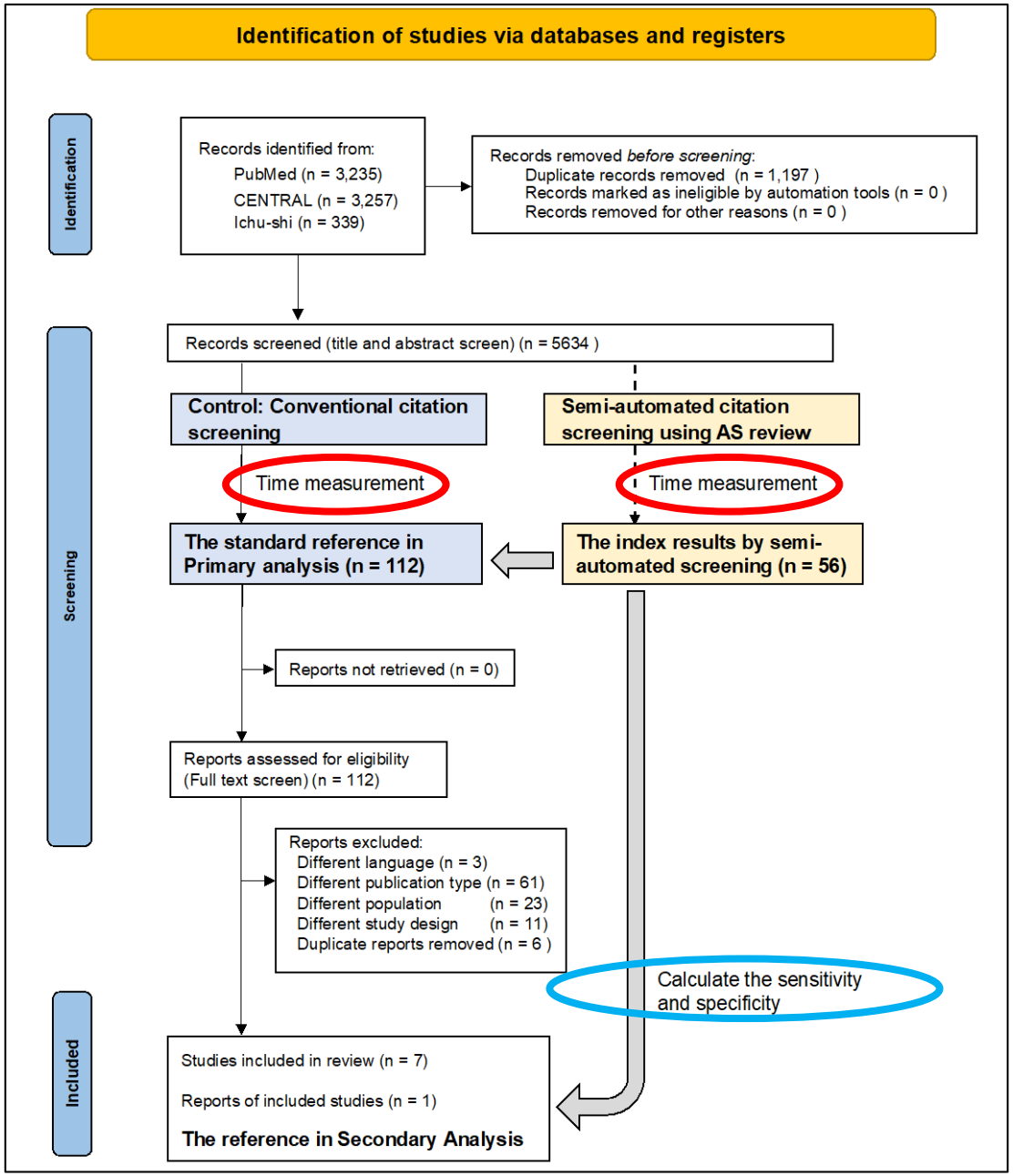


PICOに合致した文献数

スクリーニング文献数

スクリーニングにより関連のある文献が得られるプラトーに達した可能性

研究のフローチャート



1次スクリーニング

Title/abstract

2次スクリーニング

Full-text

Reviewerの背景

	手動（従来法）	自動
Reviewerの人数	18	2
年齢	38.0 (35-39)	39.5 (38-41)
男性の割合, n (%)	17 (94.4)	2 (100)
職種		
医師, n (%)	17 (94.4)	2 (100)
看護師, n (%)	1 (5.6)	0 (0)
専門領域（重複あり）		
救急, n (%)	12 (66.7)	2 (100)
集中治療, n (%)	14 (77.8)	2 (100)
麻酔, n (%)	4 (22.2)	0 (0)
その他	3 (16.7)	0 (0)
博士号, n (%)	7 (38.9)	2 (100.0)
臨床経験年数		
0-5年, n (%)	0 (0)	0 (0)
6-10年, n (%)	5 (27.8)	0 (0)
11-15年, n (%)	9 (50.0)	1 (50.0)
16-20年, n (%)	4 (22.2)	1 (50.0)
≥ 21年, n (%)	0 (0)	0 (0)
システマティックレビューの論文数		
0本, n (%)	15 (83.3)	1 (50.0)
1-5本, n (%)	3 (16.7)	0 (0)
6-10本, n (%)	0 (0)	0 (0)
≥ 11本, n (%)	0 (0)	1 (50.0)

半自動化文献スクリーニングの精度

	2次スクリーニング後の該当文献数	自動化で該当と判断した文献数	文献数
CQ1（Lactate）	17	16	4,326
CQ2（輸液制限）	8	8	2,253
CQ3（輸液）	8	7	5,634
CQ4（血圧）	4	4	3,418
CQ5（重炭酸）	4	3	1,038

※2次スクリーニング後の最終的な該当論文の識別能を評価

Oami T, et al. J Epidemiol. 2024;34(8):380-6.

文献スクリーニングの精度の比較

	手動（従来法）			自動		
	感度	特異度	陽性的中率	感度	特異度	陽性的中率
CQ1（Lactate）	1.0	0.996	0.243	0.941	0.997	0.400
CQ2（輸液制限）	1.0	0.986	0.205	1.0	0.981	0.160
CQ3（輸液）	1.0	0.987	0.071	0.875	0.994	0.125
CQ4（血圧）	1.0	0.981	0.235	1.0	0.991	0.286
CQ5（重炭酸）	1.0	0.990	0.286	0.750	0.989	0.214

※ 2次スクリーニング後の最終的な該当論文をゴールドスタンダードとした場合の感度・特異度・陽性的中率を算出した。

Oami T, et al. J Epidemiol. 2024;34(8):380-6.

文献スクリーニングの所要時間の比較

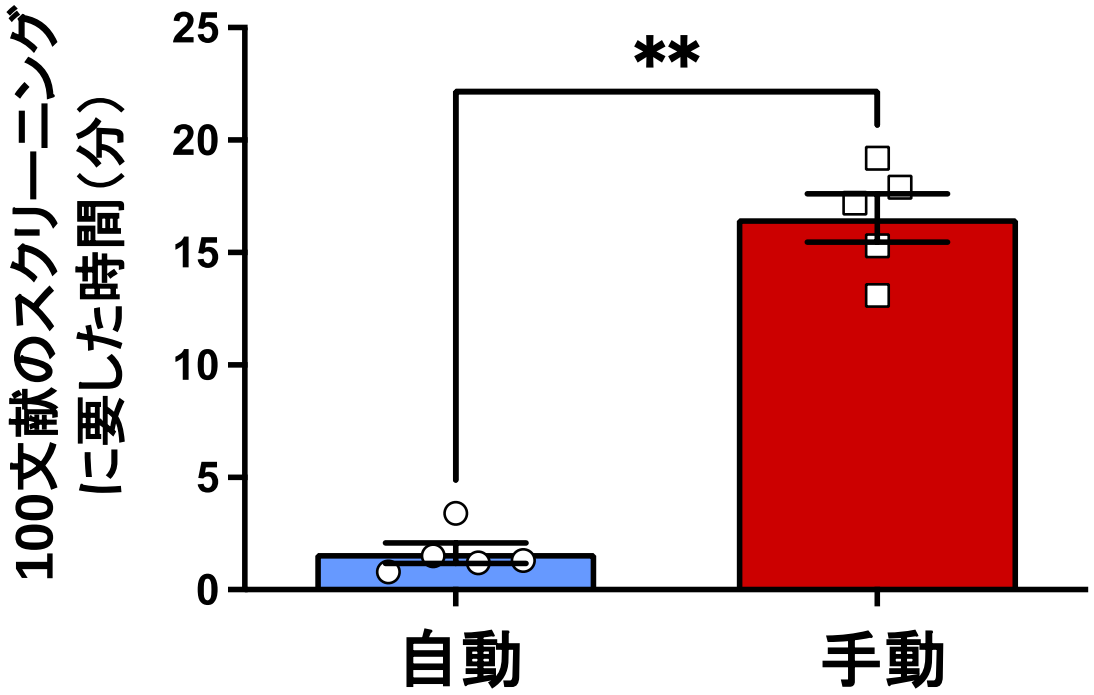
	手動（min）	自動（min）	文献数
CQ1（Lactate）	830	55	4,326
CQ2（輸液制限）	295	34	2,253
CQ3（輸液）	861	43	5,634
CQ4（血圧）	589	41	3,418
CQ5（重炭酸）	186	35	1,038

※複数人で分担して作業を行った場合には分担範囲の平均所要時間をそれぞれ合算した時間を所要時間とした。

Oami T, et al. J Epidemiol. 2024;34(8):380-6.

1CQごとの文献スクリーニングの所要時間の比較

** p < 0.01



※異なるCQ間で比較するために1CQあたりの所要時間を総文献数で割った値を算出した。

Oami T, et al. J Epidemiol. 2024;34(8):380-6.

結果のまとめ

- ✓ 最終的な該当論文をゴールドスタンダードとした場合の半自動化ソフトの感度は0.75～1.0，特異度は0.98～0.99，陽性的中率は0.13～0.40（従来法：0.07～0.29）であった。

- ✓ 半自動化ソフトを用いた文献スクリーニングの作業時間は従来の方法に比べて約90%短縮した。

文献スクリーニング支援ソフトの精度に関する過去の研究との比較

- ✓ 文献スクリーニング支援ソフトであるAbstrackrの感度は0.75であった。

  Gates A, et al: Syst Rev. 2019;7(1):45.

- ✓ 別の自然言語処理を使った文献スクリーニング支援ソフトCovidenceの感度は0.90～0.92であった。

  Perlman-Arrow S, et al: Res Synth Methods. 2023;14(4):608-21.

本研究の感度は0.75～1.0であり，別の文献スクリーニング支援ソフトを用いた過去の報告と遜色ない精度であった。

作業負担に関する過去の研究との比較

✓ 文献スクリーニング支援ソフトCovidenceを用いた研究では作業時間が45%短縮したと報告されている。

Perlman-Arrow S, et al: Res Synth Methods. 2023;14(4):608-21.

今回の研究では通常の方法と比べて作業時間が90%短縮しており，作業負担軽減の観点からは自動化ソフトの使用は有用である可能性がある。

※今回用いた自動化文献スクリーニングソフトの作業時間は，学習データとしてどれだけ多くの文献レビューを実行するかに依存しており，精度向上のために文献レビュー数を増やすことによって作業時間は増加することになる。

最終組み入れ文献の見逃しについて

最終組み入れ文献のうち，自動化ソフトで関連なしと判断された文献リスト

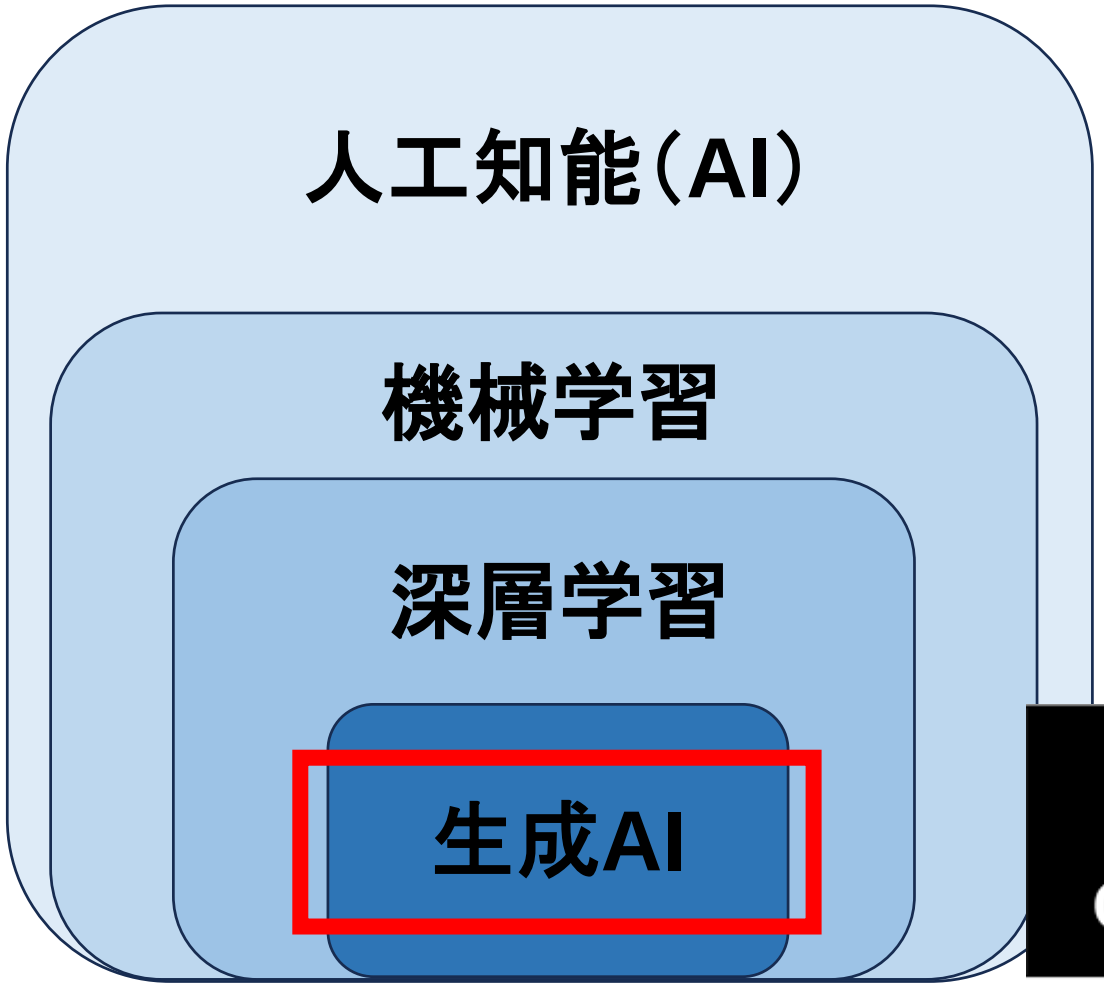
	Year	Journal	Title
CQ1	2019	American Journal of Respiratory and Critical Care Medicine	Balanced crystalloids versus saline in sepsis. A secondary analysis of the SMART Clinical Trial
CQ3	2023	Critical Care Medicine	Long-term outcome of severe metabolic acidemia in ICU patients, a BICAR-ICU trial post hoc analysis
CQ4	1995	New England Journal of Medicine	A trial of goal-oriented hemodynamic therapy in critically ill patients. SvO_2 Collaborative Group

半自動化文献スクリーニングソフトによる見逃しを回避する方法

- 適切なキー論文の検索
- スクリーニング文献数の追加による学習データの充足
- 従来のスクリーニングと半自動化ソフトを用いたハイブリッド法

生成AIを用いた
文献スクリーニング

AIの分類



人間のように学習・判断・推論する能力を持つコンピュータシステムの総称

データからパターンを学習して予測や分類を行う技術

人間の脳にヒントを得たニューラルネットワークを使って複雑な特徴を自動で学習する方法

OpenAI　Claude　Gemini

診療ガイドライン作成のための生成AIを用いた文献スクリーニングの精度と作業負荷軽減の検証

目的

日本版敗血症診療ガイドラインの作成において，
生成AIを用いた文献スクリーニングの精度と作業
負荷軽減の評価を行うこと。

GPT-4 Turbo

生成AIを用いた文献スクリーニング

You

プロンプト

You are conducting a systematic review and meta-analysis, focusing on a specific area of medical research. Your task is to evaluate research studies and determine whether they should be included in your review. To do this, each study must meet the following criteria:

Target Patients: Adult patients (18 years old or older) diagnosed with or suspected of having infection, bacteremia, or sepsis.
Intervention: The study investigates the effects of balanced crystalloid administration.
Comparison: The study compares the above intervention with 0.9% sodium chloride administration.
Study Design: The study must be a randomized controlled trial.
Additionally, any study protocol that meets these criteria should also be included.

However, you should exclude studies in the following cases:

The study does not meet all of the above eligibility criteria.
The study's design is not a randomized controlled trial. Examples of unacceptable designs include case reports, observational studies, systematic reviews, review articles, animal experiments, letters to editors, and textbooks.
After reading the title and abstract of a study, you will decide whether to include or exclude it based on these criteria. Please answer with include or exclude only.

Title: ----------

Abstract

ChatGPT

Include

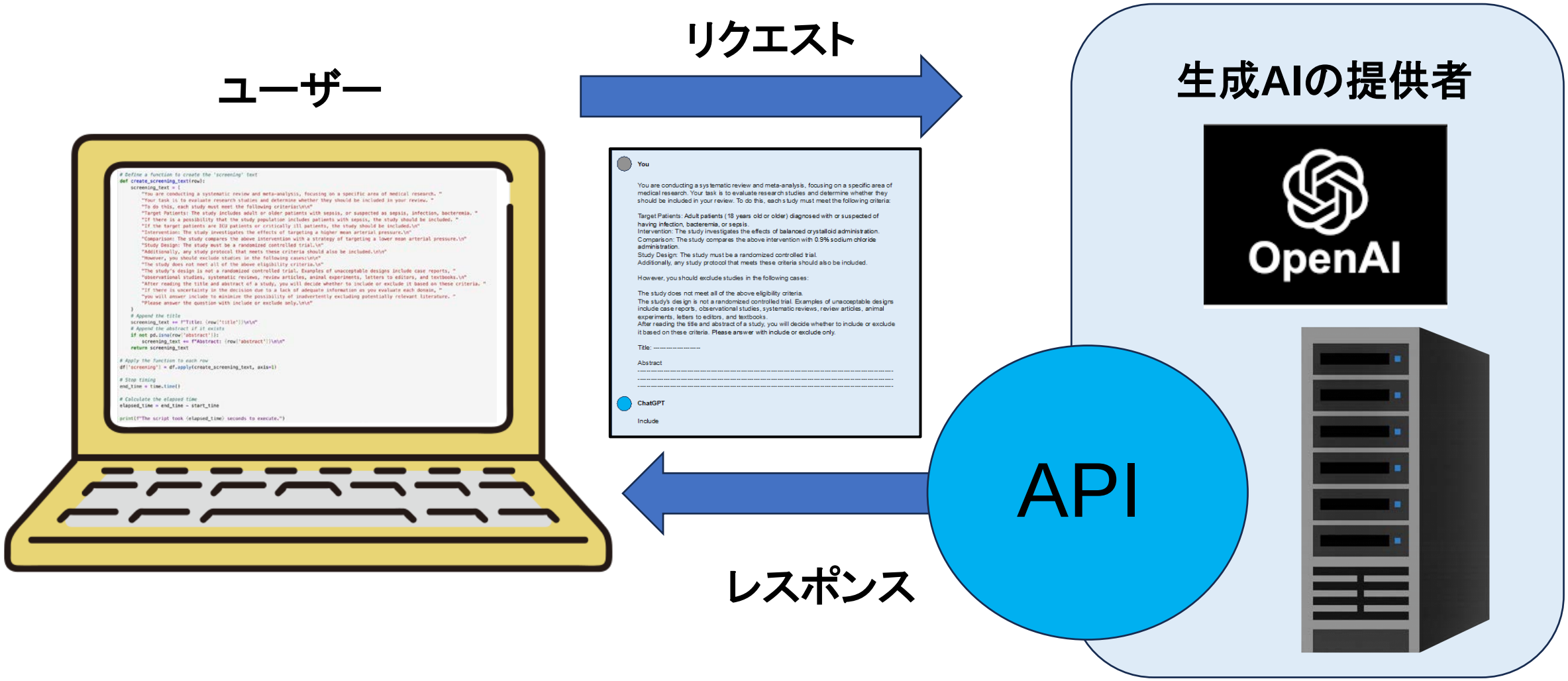
役割を与える

PICOを明示する

作業内容を指示する

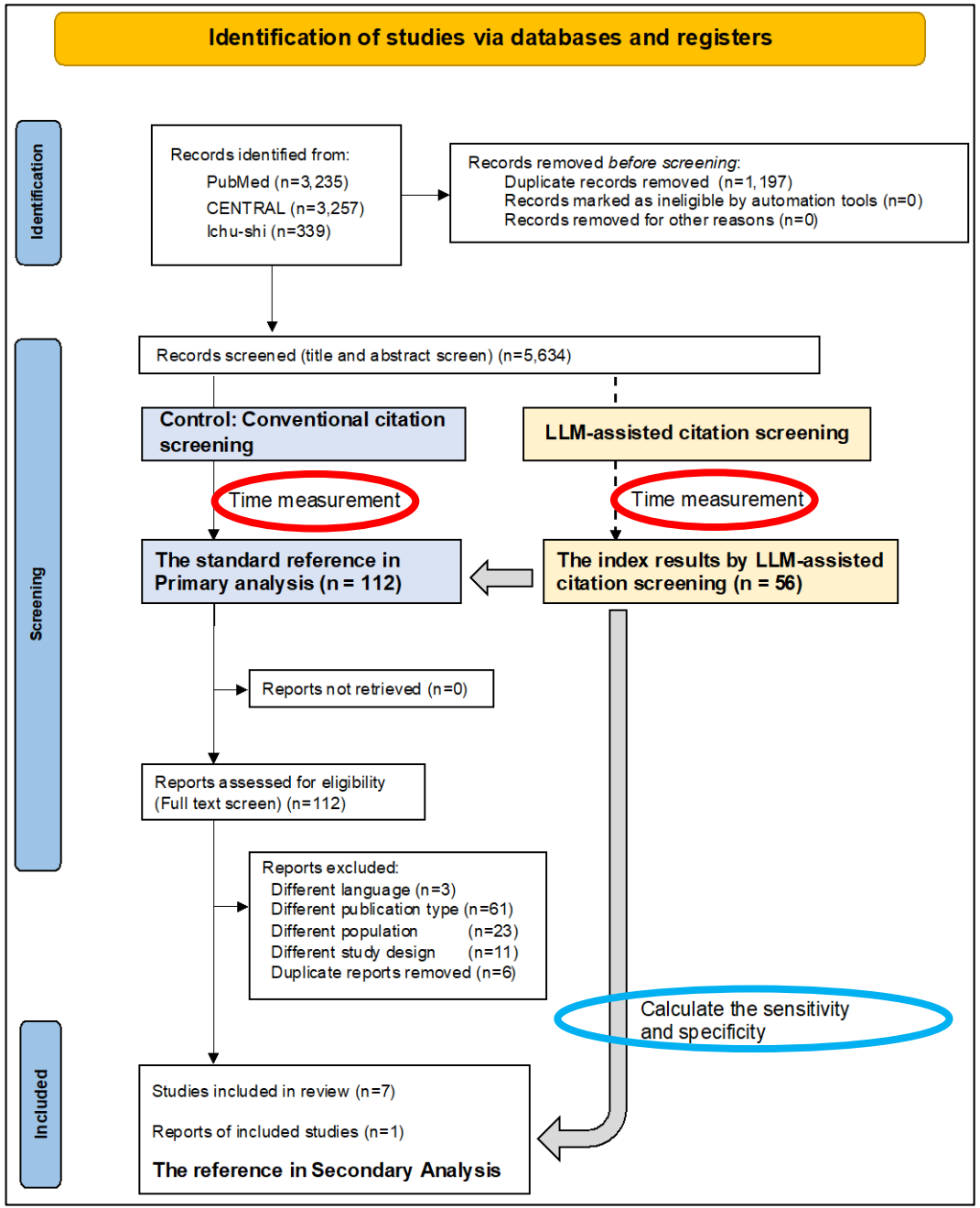
文献情報を示す

生成AIを用いた文献スクリーニングの実際



API: application programming interface

研究のフローチャート



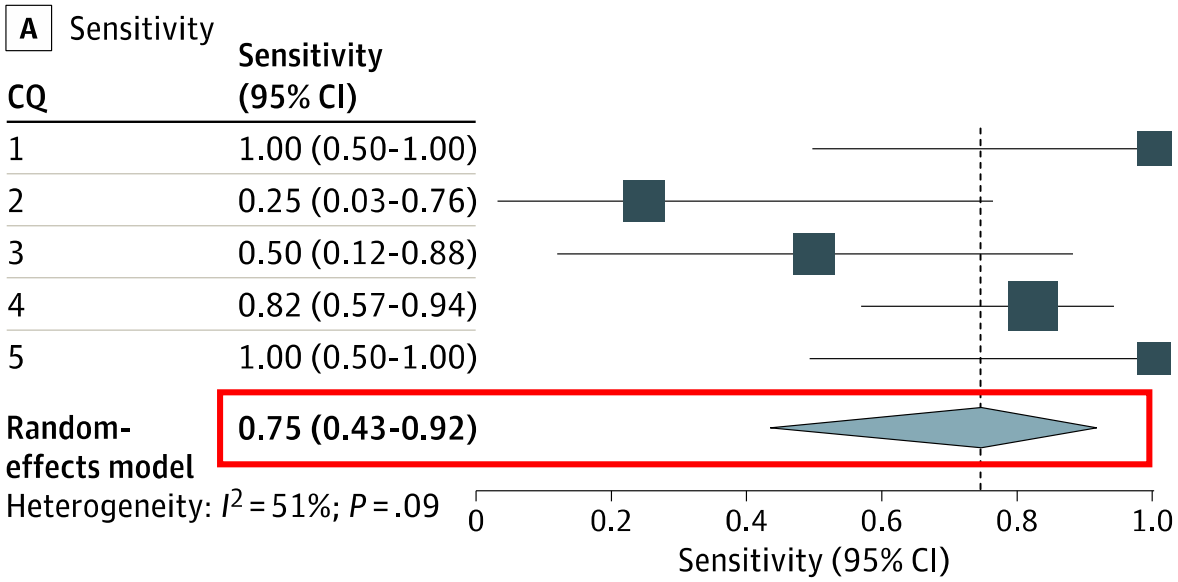
1次スクリーニング

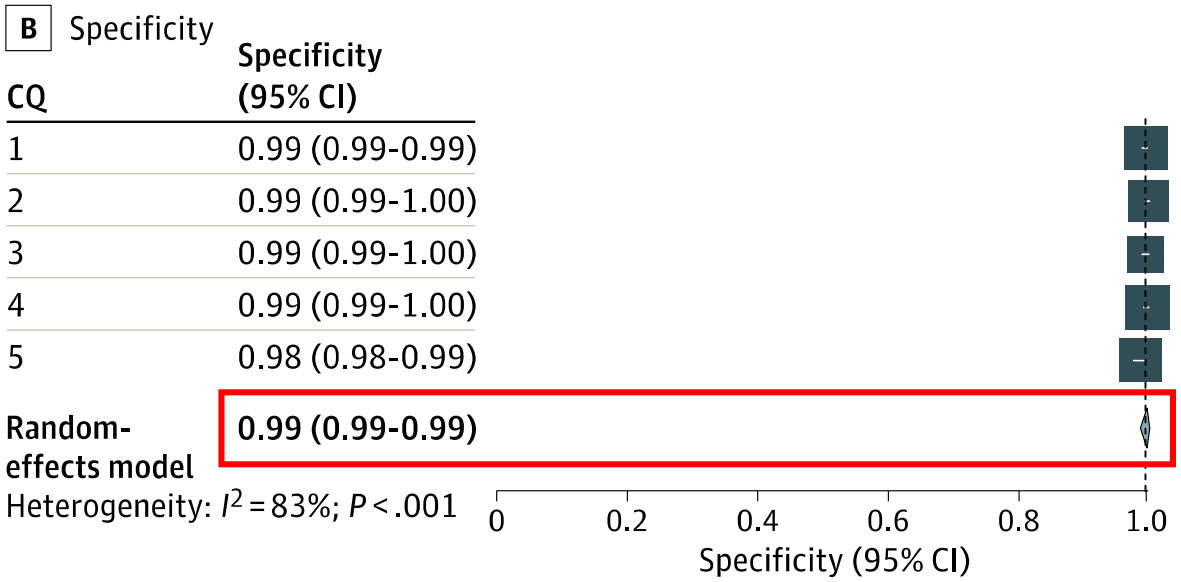
Title/abstract

2次スクリーニング

Full-text

生成AIを用いた文献スクリーニングの精度





※2次スクリーニング後の最終的な該当論文を使用

Oami T, et al. JAMA Netw Open. 2024;7(7):e2420496.

修正プロンプト

You

You are conducting a systematic review and meta-analysis, focusing on a specific area of medical research. Your task is to evaluate research studies and determine whether they should be included in your review. To do this, each study must meet the following criteria:

Target Patients: The study includes adult patients diagnosed with or suspected of having infection, bacteremia, or sepsis. If there is a possibility that the study population includes patients with sepsis, the study should be included.
Intervention: The study investigates the effects of balanced crystalloid administration.
Comparison: The study compares the above intervention with 0.9% sodium chloride administration.
Study Design: The study must be a randomized controlled trial.
Additionally, any study protocol that meets these criteria should also be included.

However, you should exclude studies in the following cases:

The study does not meet all of the above eligibility criteria.
The study's design is not a randomized controlled trial. Examples of unacceptable designs include case reports, observational studies, systematic reviews, review articles, animal experiments, letters to editors, and textbooks.
After reading the title and abstract of a study, you will decide whether to include or exclude it based on these criteria. If there is uncertainty in the decision due to a lack of adequate information as you evaluate each domain, you will answer include to minimize the possibility of inadvertently excluding potentially relevant literature. Please answer with include or exclude only.

Title: ----------

Abstract

ChatGPT

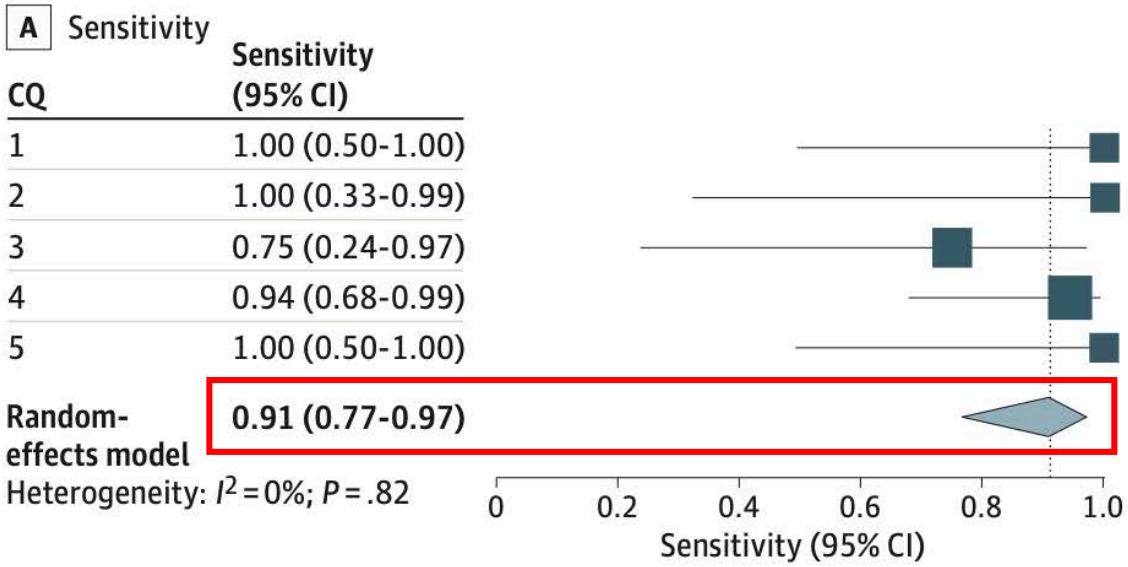
Include

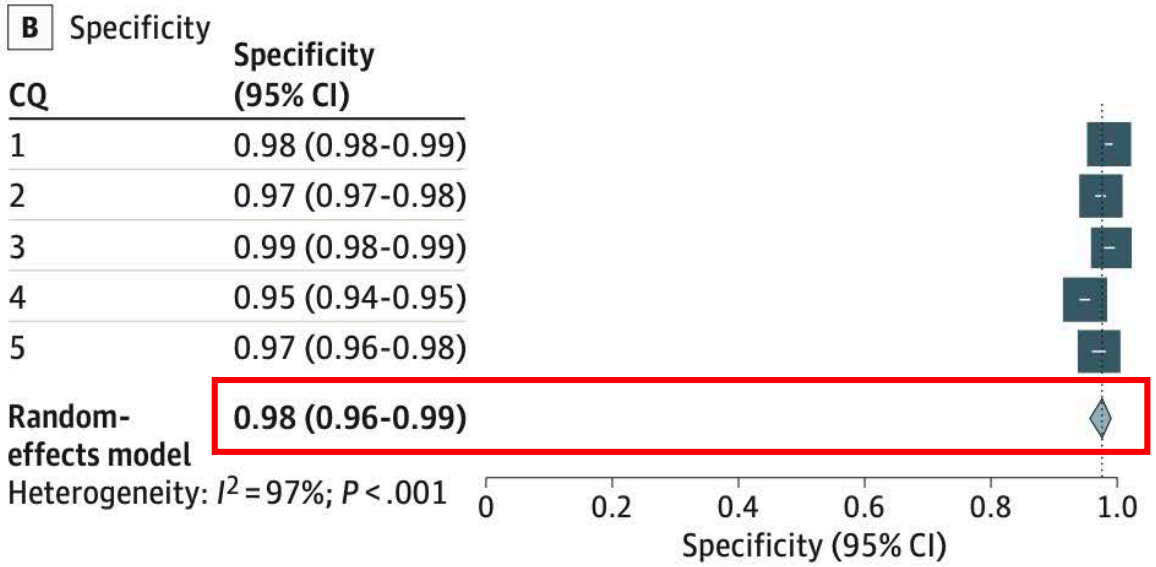
] Populationの取りこぼしを防ぐプロンプトを追加

] 判断が難しい場合にはIncludeを選択させるプロンプトを追加

感度の上昇につながる可能性

修正プロンプトを用いた生成AIによる文献スクリーニングの精度



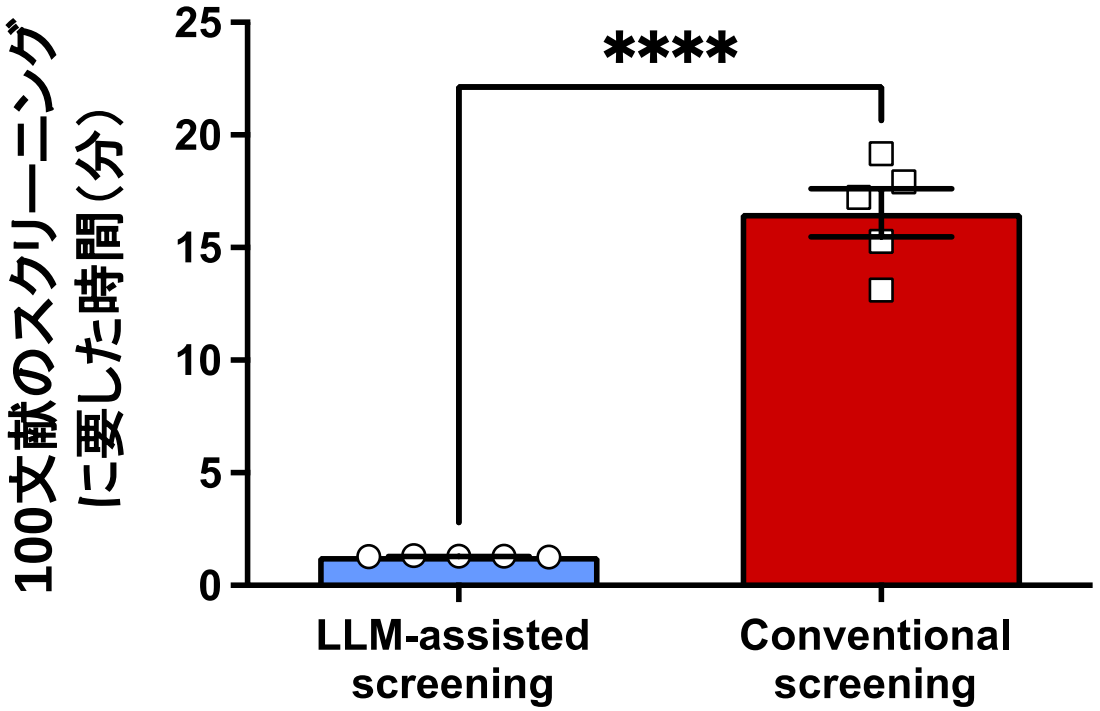


※2次スクリーニング後の最終的な該当論文を使用

Oami T, et al. JAMA Netw Open. 2024;7(7):e2420496.

1CQごとの文献スクリーニングの所要時間の比較

** p < 0.01



※異なるCQ間で比較するために1CQあたりの所要時間を総文献数で割った値を算出した。

Oami T, et al. JAMA Netw Open. 2024;7(7):e2420496.

結果のまとめ

- ✓ 最終的な該当論文をゴールドスタンダードとした場合の生成AIを用いた文献スクリーニングの統合感度は0.75 (0.25～1.00), 統合特異度 0.99 (0.98～0.99)であった。

- ✓ 修正プロンプトを用いた場合の統合感度は0.91 (0.75～1.0), 統合特異度 0.98 (0.95～0.99)であった。

- ✓ 生成AIを用いた文献スクリーニングの作業時間は従来の方法に比べて約90%短縮した。

生成AIを用いた文献スクリーニングの精度に関する研究

- ✓ 過去の研究を統合したメタアナリシスでは感度は0.73 (95% CI: 0.57–0.85)，特異度は0.99 (95% CI: 0.97–0.99)と報告されている。

- ✓ 修正プロンプトを用いることで感度0.98 (95% CI: 0.74–1.00)へ上昇した一方，特異度0.98 (95% CI: 0.94–0.99)と維持された。

Dai Z-Y, et al. SSRN Scholarly Paper; 2024.

1次スクリーニング前のプレスクリーニングやセカンドレビューワーとして生成AIを活用することで，文献スクリーニング作業の効率化が期待される。

生成AIを用いた文献スクリーニングの精度向上に向けて

- **最新の生成AIモデルを用いる**

  GPT-3.5よりもGPT-4の方が文献スクリーニングの精度が向上

  Oami T, et al. JMIR Med Inform. 2025;13:e64682.

- **プロンプトエンジニアリング**

  Few-shot prompting, chain-of-thought strategy, majority-vote strategy

- **検索拡張生成 (RAG)の活用**

  専門的な学習データの追加

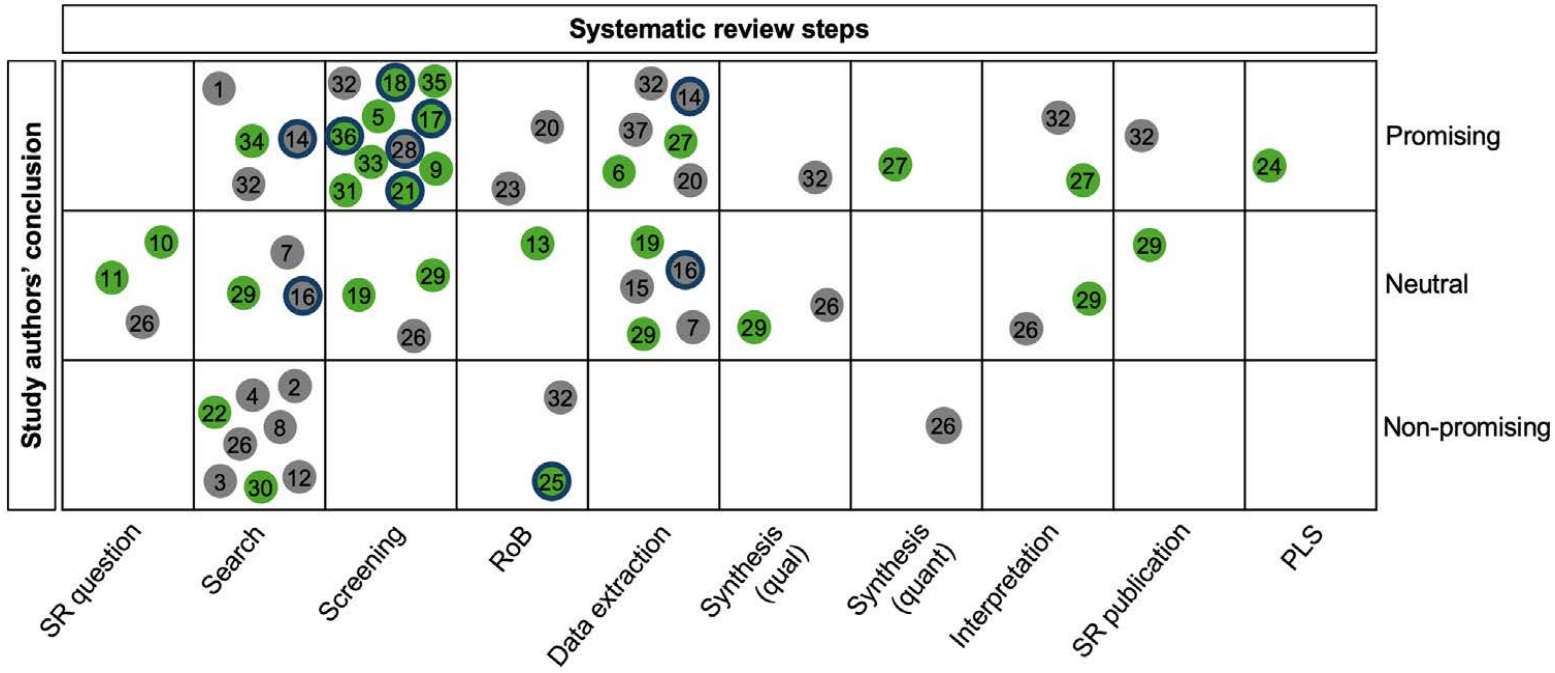
将来的に生成AIを用いて文献スクリーニング作業を完全に自動化することができる可能性あり

AIを用いたガイドライン作成の展望

生成AIを用いたsystematic reviewの自動化

SRの工程	研究結果の概要	課題や限界
文献検索式の作成	GPT-4によりPICOに基づいたPubMed検索式の自動生成が可能であった。	同義語やMeSH termの不足 専門用語の精度不足
文献スクリーニング	タイトルと抄録を用いたスクリーニングにおいて高い精度と90%以上の作業時間の短縮が得られた。	研究分野による精度の変動 モデルやプロンプトによる精度の差
データ抽出	論文の表や本文から自動的に高い精度で数値やアウトカム情報を抽出することが可能であった。	設定条件やプロンプトへの精度の依存 Hallucinationのリスク
リスクオブバイアス評価	生成AIによる自動判定は，一定の条件下で人間の評価者とほぼ同等のパフォーマンスを示した。	特定の領域での精度の低下 プロンプトによる精度への影響の大きさ

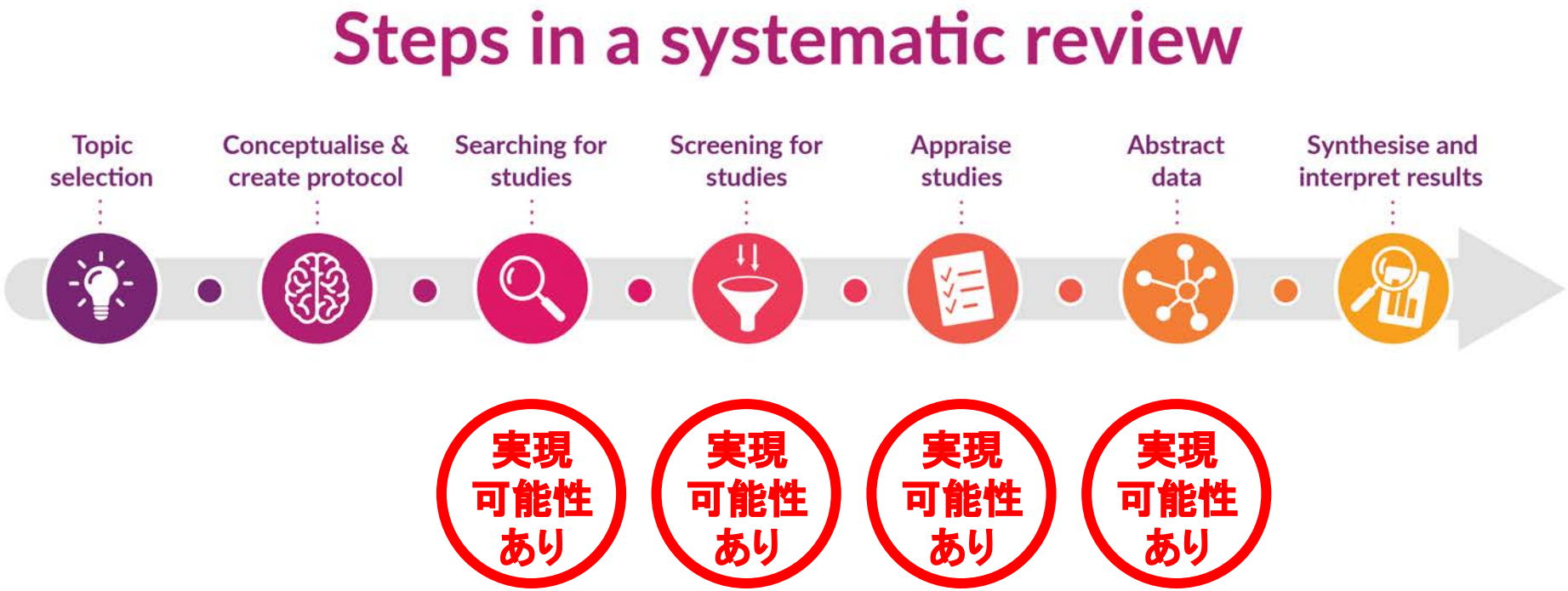
生成AIを用いたsystematic reviewの自動化



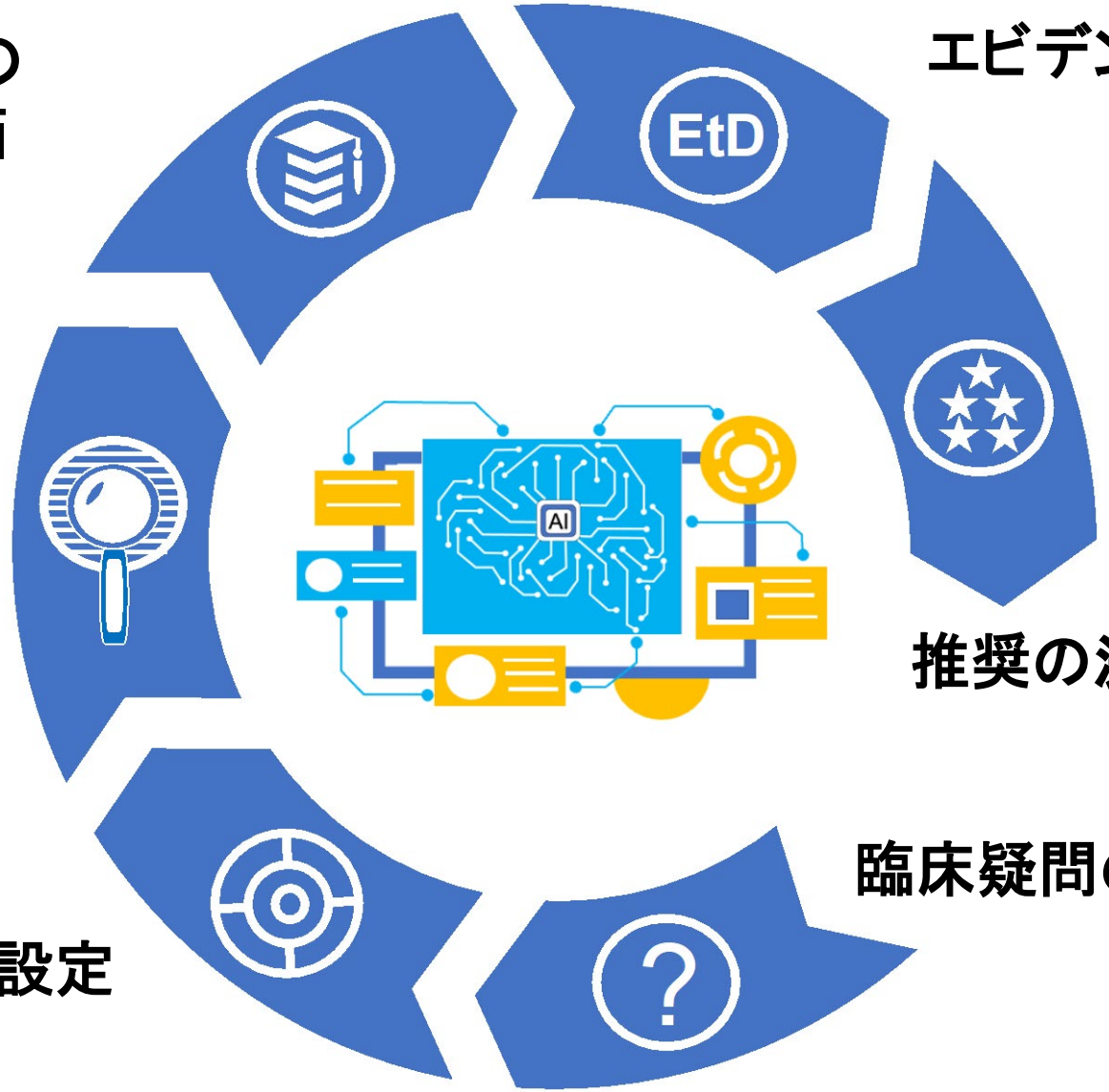
Lieberum JL, et al. J Clin Epidemiol. 2025;181:111746.

システマティックレビューの工程によって生成AIを用いた自動化の作業の質や精度が異なる可能性

AIを用いたsystematic reviewの自動化



AIを用いた診療ガイドライン作成の自動化



エビデンスの確実性評価

エビデンスから推奨へ

システマティックレビュー

推奨の決定

アウトカムの設定

臨床疑問の立案

AIを用いたガイドライン作成における課題

- Hallucinationや誤情報による質の低下
- 人間が担うべき作業の明確化
- 情報のコンタミネーションリスク

結 語

1. 半自動化スクリーニングソフトは文献スクリーニングの作業時間を短縮し，診療ガイドラインを作成するための重要な研究を同定した。

2. 生成AIを用いた文献スクリーニングはプロンプトの修正により精度が向上し，許容範囲内の感度と高い特異度を示した。

3. AIを用いたガイドライン作成の有用性を検証するために，さらなる研究が必要である。

謝辞

本講演の発表内容に関して，ご指導･ご協力頂いた先生方に感謝を申し上げます。

日本版敗血症診療ガイドライン2024特別委員会

中田 孝明，志馬 伸朗，櫻谷 正明，福田 龍将，柏浦 正広，
山本 良平，佐藤 威仁，松浦 裕司，彦根 麻由，山田 浩平，
湯本 哲也，鉄原 健一，長澤 宏樹，米倉 寛，藤永 潤，
久宗 遼，木庭 茂，野浪 豪，恒光 健史，　濵井 康貴，
若林 侑起，水野 彰人，雨宮 優，村田 哲平（敬称略）

ご清聴ありがとうございました！