Global Evidence Summit

ebq
Evidence-based quality

Clinical Guidelines Resources

Leveraging off-the-shelf AI to facilitate title/ abstract screening, full-text review, and data extraction

Thomas Schofield, Chris Wolfkiel, & Jon Heald

Using evidence. Improving lives.

Declaration of Conflict of interest

To the best of my knowledge, I declare that neither I nor any of my co-authors/co-presenters, not any of my/our close family members, have had employment, received research support or other funding from, or had any other professional relationship with an entity directly involved in the production, manufacture, distribution or sale of tobacco, tobacco products, weapons or arms, or have represented the interests of any such entities in any way.

TS and CH have no actual or potential conflict of interest in relation to this presentation. CW is a consultant to Indico Solutions.

Global Evidence Summit

PRE-WORKSHOP SURVEY RESULTS

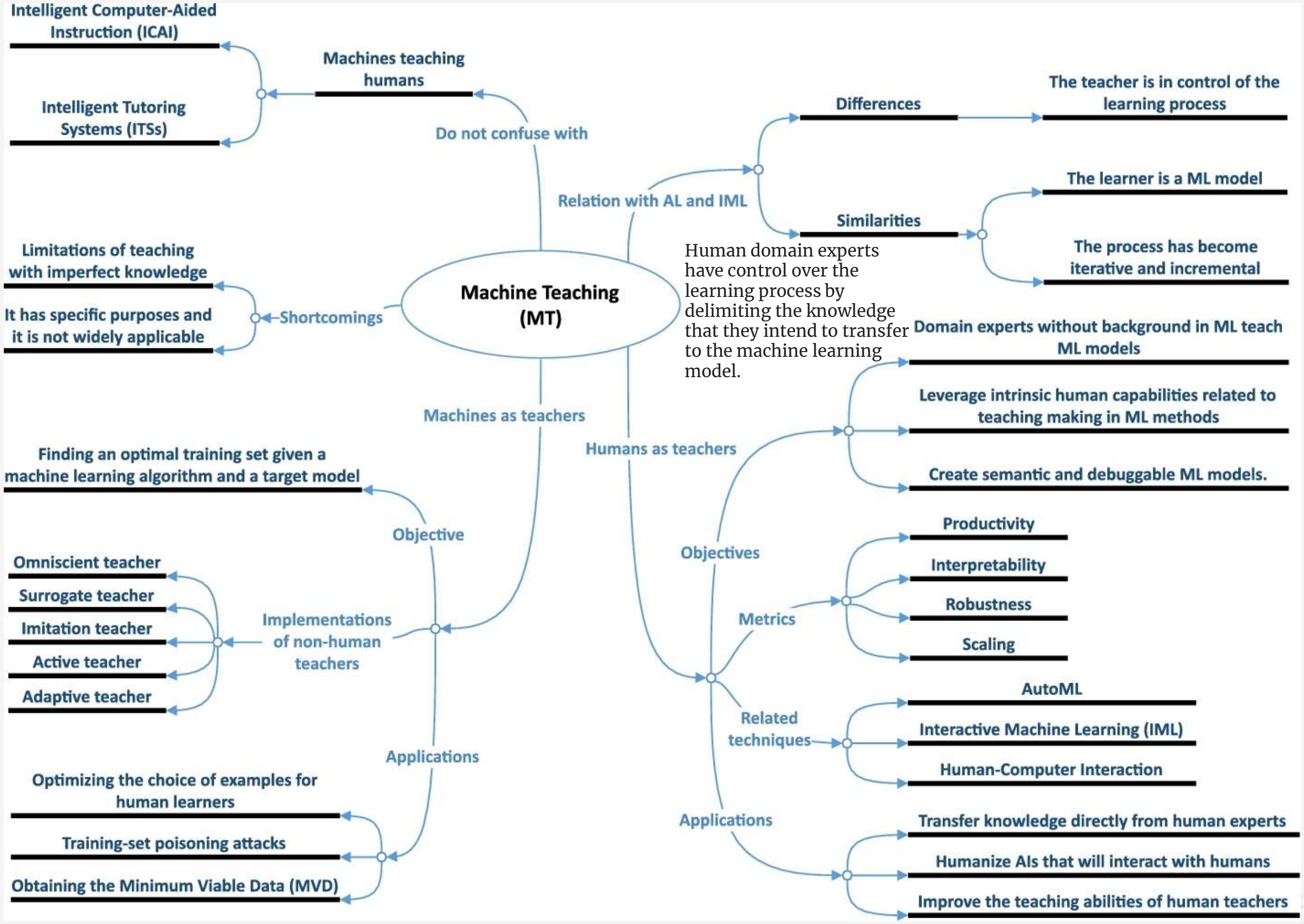
Global Evidence Summit

Systematic reviews: Artificial intelligence helps?

SR Workflow Options (using TIAB screening as an example):

- **Humans do it alone** Dual independent screening, with discrepancies adjudicated by a third reviewer **(typical)**
- **Human and machine work in parallel** One human screens, and one Gen AI screens **(today's focus)**
- **Human and machine work together** Human trains AI/ML, which helps human screen faster **(emerging/ongoing)**
- **Machines do it alone** Two different AI conduct screening **(someday?)**

Figure taken from Mosqueira-Rey et al. 2022.


Machine Teaching (MT)
Do not confuse with
Machines teaching humans
Intelligent Computer-Aided Instruction (ICAI)
Intelligent Tutoring Systems (ITSs)
Shortcomings
Limitations of teaching with imperfect knowledge
It has specific purposes and it is not widely applicable
Machines as teachers
Objective
Finding an optimal training set given a machine learning algorithm and a target model
Implementations of non-human teachers
Omniscient teacher
Surrogate teacher
Imitation teacher
Active teacher
Adaptive teacher
Applications
Optimizing the choice of examples for human learners
Training-set poisoning attacks
Obtaining the Minimum Viable Data (MVD)
Relation with AL and IML
Differences
The teacher is in control of the learning process
Similarities
The learner is a ML model
The process has become iterative and incremental
Humans as teachers
Objectives
Domain experts without background in ML teach ML models
Leverage intrinsic human capabilities related to teaching making in ML methods
Create semantic and debuggable ML models.
Metrics
Productivity
Interpretability
Robustness
Scaling
Related techniques
AutoML
Interactive Machine Learning (IML)
Human-Computer Interaction
Applications
Transfer knowledge directly from human experts
Humanize AIs that will interact with humans
Improve the teaching abilities of human teachers

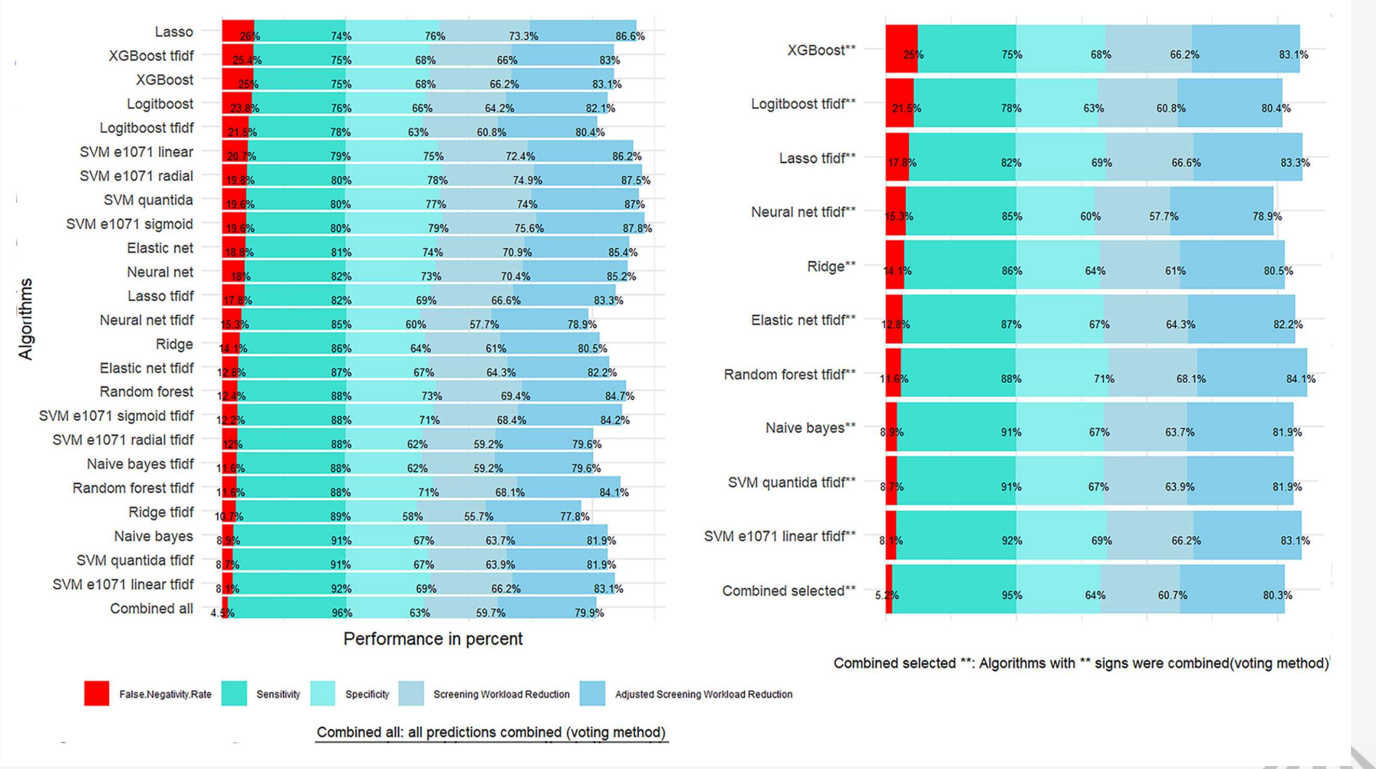
Human domain experts have control over the learning process by delimiting the knowledge that they intend to transfer to the machine learning model.

Systematic reviews: Artificial intelligence helps!

Observed Benefits in:

- **Speed**
  - Fabiano et al., 2024
  - Pijls, 2024
  - van Dijk et al., 2023
- **Accuracy**
  - Belur et al 2018
  - Beresford et al. 2022
  - Hanegraaf et al., 2024
- **Project Effiencies**

Efficiency of AI-led screening



Combined selected **: Algorithms with ** signs were combined(voting method)

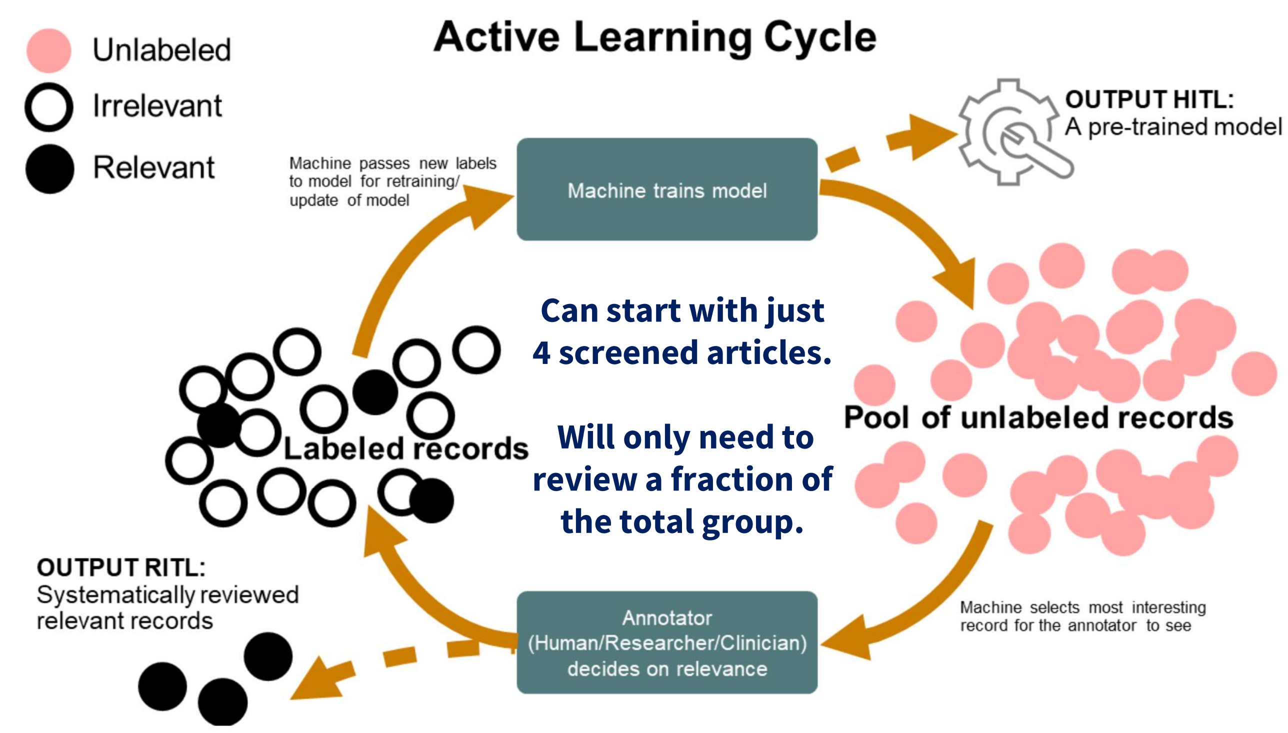
Combined all: all predictions combined (voting method)

Figure taken from Kebede et al., 2022.

Human and AI/ML working together: An example

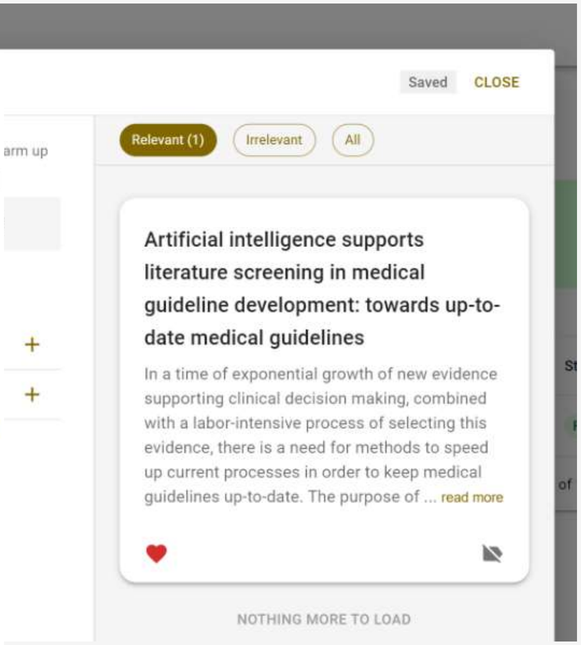
With ASReview LAB humans decide which articles to accept, the AI only sorts _the order in which the abstracts are reviewed_

2 minute intro


Active Learning Cycle
Unlabeled
Irrelevant
Relevant
Machine passes new labels to model for retraining/ update of model
Machine trains model
OUTPUT HITL:
A pre-trained model
Labeled records
Pool of unlabeled records
Machine selects most interesting record for the annotator to see
Annotator (Human/Researcher/Clinician) decides on relevance
OUTPUT RITL:
Systematically reviewed relevant records

Can start with just 4 screened articles.

Will only need to review a fraction of the total group.



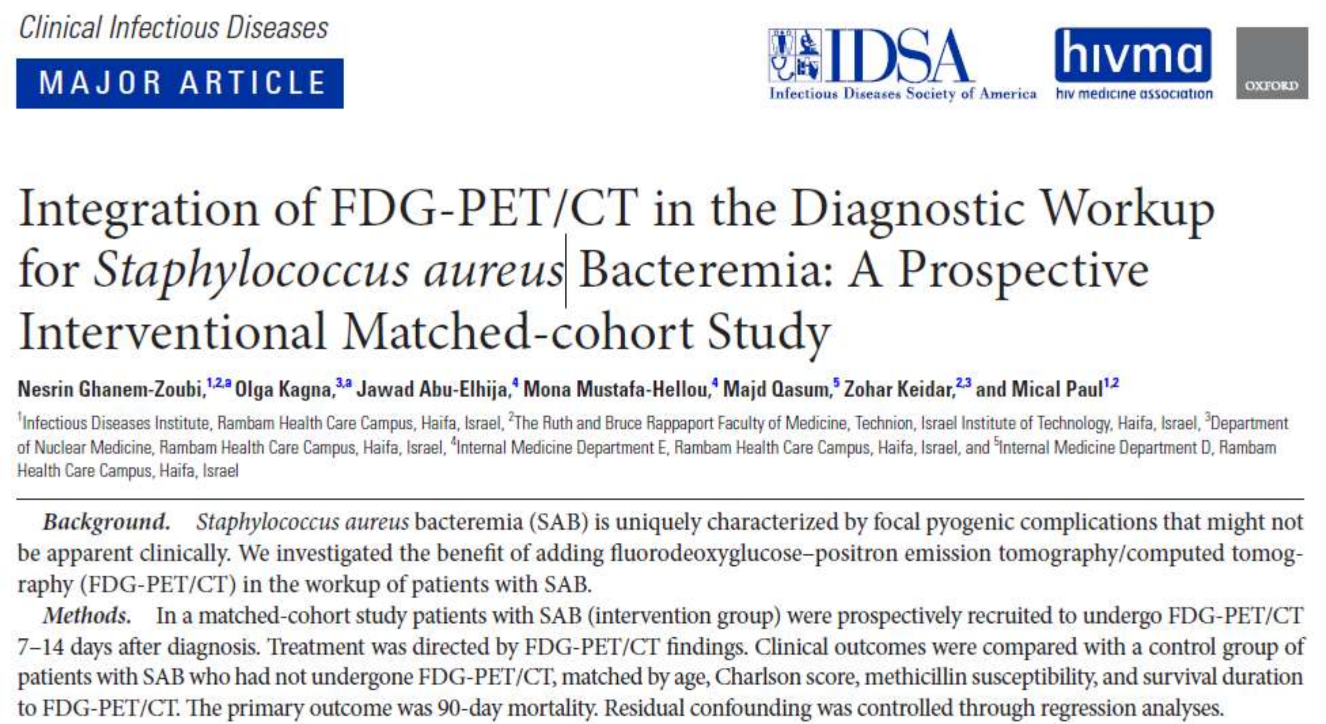
Artificial intelligence supports literature screening in medical guideline development: towards up-to-date medical guidelines

In a time of exponential growth of new evidence supporting clinical decision making, combined with a labor-intensive process of selecting this evidence, there is a need for methods to speed up current processes in order to keep medical guidelines up-to-date. The purpose of ... read more

Figure taken from https://asreview.nl/blog/active-learning-explained/

Human and Machine working in parallel: Example 1

Ghanem-Zoubi et al., 2020



Clinical Infectious Diseases

MAJOR ARTICLE

IDSA Infectious Diseases Society of America

hivma hiv medicine association

OXFORD

Integration of FDG-PET/CT in the Diagnostic Workup for *Staphylococcus aureus* Bacteremia: A Prospective Interventional Matched-cohort Study

Nesrin Ghanem-Zoubi,[1,2,a] Olga Kagna,[3,a] Jawad Abu-Elhija,[4] Mona Mustafa-Hellou,[4] Majd Qasum,[5] Zohar Keidar,[2,3] and Mical Paul[1,2]

[1]Infectious Diseases Institute, Rambam Health Care Campus, Haifa, Israel, [2]The Ruth and Bruce Rappaport Faculty of Medicine, Technion, Israel Institute of Technology, Haifa, Israel, [3]Department of Nuclear Medicine, Rambam Health Care Campus, Haifa, Israel, [4]Internal Medicine Department E, Rambam Health Care Campus, Haifa, Israel, and [5]Internal Medicine Department D, Rambam Health Care Campus, Haifa, Israel

---

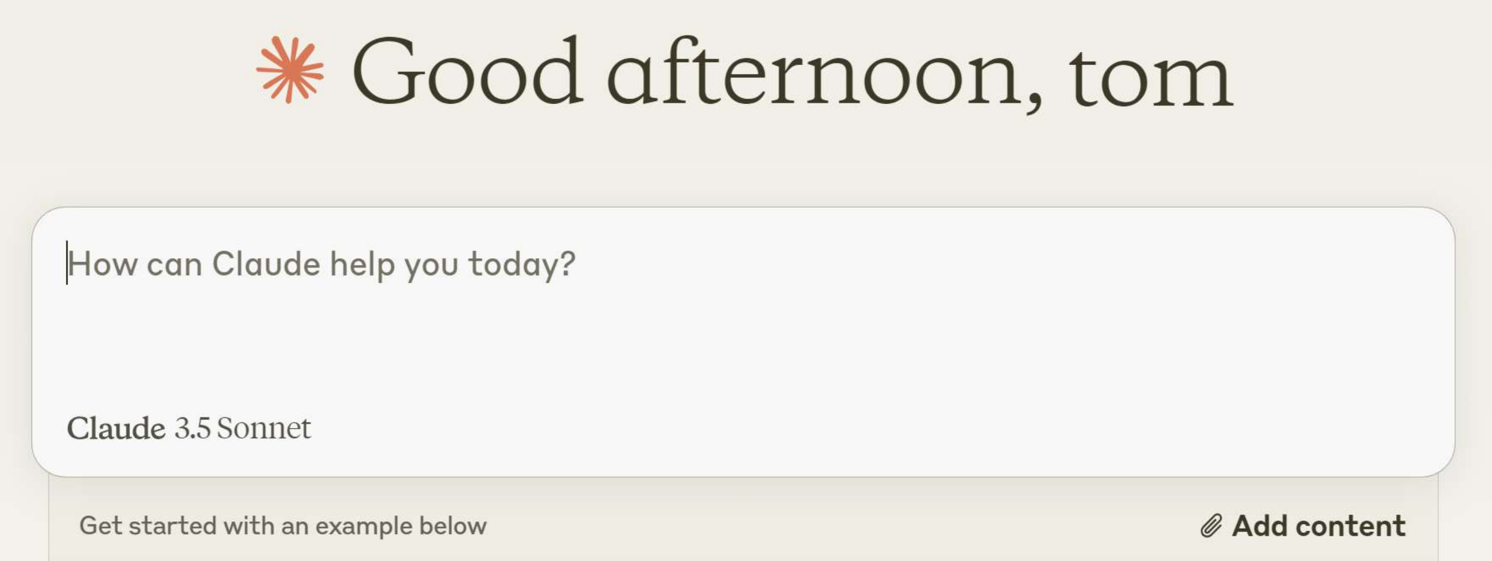
Background. *Staphylococcus aureus* bacteremia (SAB) is uniquely characterized by focal pyogenic complications that might not be apparent clinically. We investigated the benefit of adding fluorodeoxyglucose–positron emission tomography/computed tomography (FDG-PET/CT) in the workup of patients with SAB.

Methods. In a matched-cohort study patients with SAB (intervention group) were prospectively recruited to undergo FDG-PET/CT 7–14 days after diagnosis. Treatment was directed by FDG-PET/CT findings. Clinical outcomes were compared with a control group of patients with SAB who had not undergone FDG-PET/CT, matched by age, Charlson score, methicillin susceptibility, and survival duration to FDG-PET/CT. The primary outcome was 90-day mortality. Residual confounding was controlled through regression analyses.

https://tinyurl.com/269t7z2d

Today's Workshop: Human and machine working in parallel

To begin, please navigate to the following link: https://claude.ai



Good afternoon, tom

How can Claude help you today?

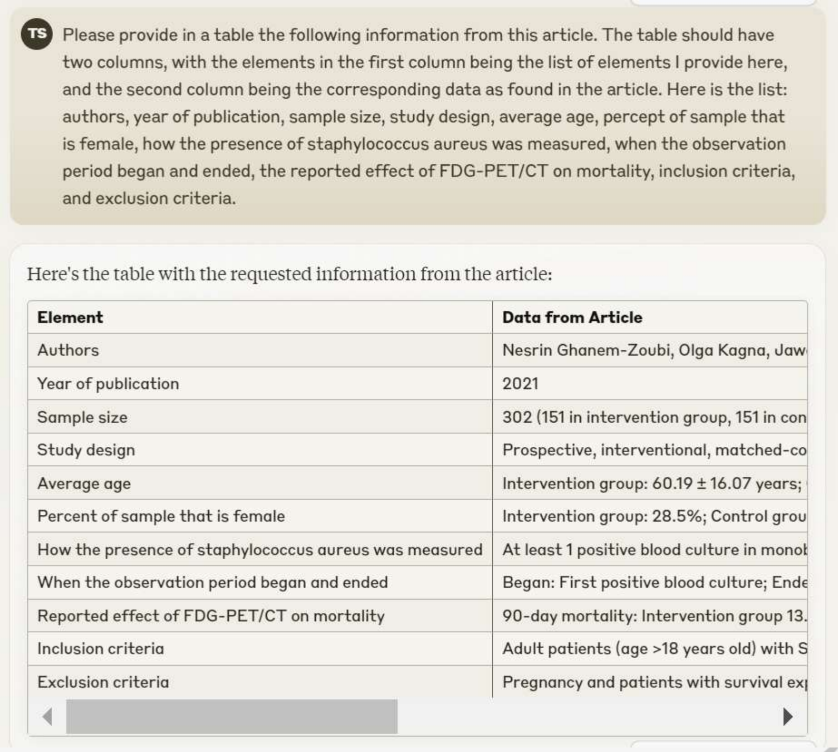
These exercises can also be tried with ChatGPT at https://openai.com/chatgpt/

Human and machine working in parallel : Extraction

Example prompt: Ghanem-Zoubi et al.

Please provide in a table the following information from this article. The table should have two columns, with the elements in the first column being the list of elements I provide here, and the second column being the corresponding data as found in the article. Here is the list: authors, year of publication, sample size, study design, inclusion criteria, and exclusion criteria, average age, percent of sample that is female, *how the presence of staphylococcus aureus was measured, when the observation period began and ended, the reported effect of FDG-PET/CT on mortality*.

Claude's response:



Please provide in a table the following information from this article. The table should have two columns, with the elements in the first column being the list of elements I provide here, and the second column being the corresponding data as found in the article. Here is the list: authors, year of publication, sample size, study design, average age, percept of sample that is female, how the presence of staphylococcus aureus was measured, when the observation period began and ended, the reported effect of FDG-PET/CT on mortality, inclusion criteria, and exclusion criteria.

Here's the table with the requested information from the article:

Element	**Data from Article**
Authors	Nesrin Ghanem-Zoubi, Olga Kagna, Jaw
Year of publication	2021
Sample size	302 (151 in intervention group, 151 in con
Study design	Prospective, interventional, matched-co
Average age	Intervention group: 60.19 ± 16.07 years;
Percent of sample that is female	Intervention group: 28.5%; Control grou
How the presence of staphylococcus aureus was measured	At least 1 positive blood culture in monob
When the observation period began and ended	Began: First positive blood culture; Ende
Reported effect of FDG-PET/CT on mortality	90-day mortality: Intervention group 13.
Inclusion criteria	Adult patients (age >18 years old) with S
Exclusion criteria	Pregnancy and patients with survival ex

Prompt adapted from Alshami et al., 2023

Human and machine working in parallel: FT Review

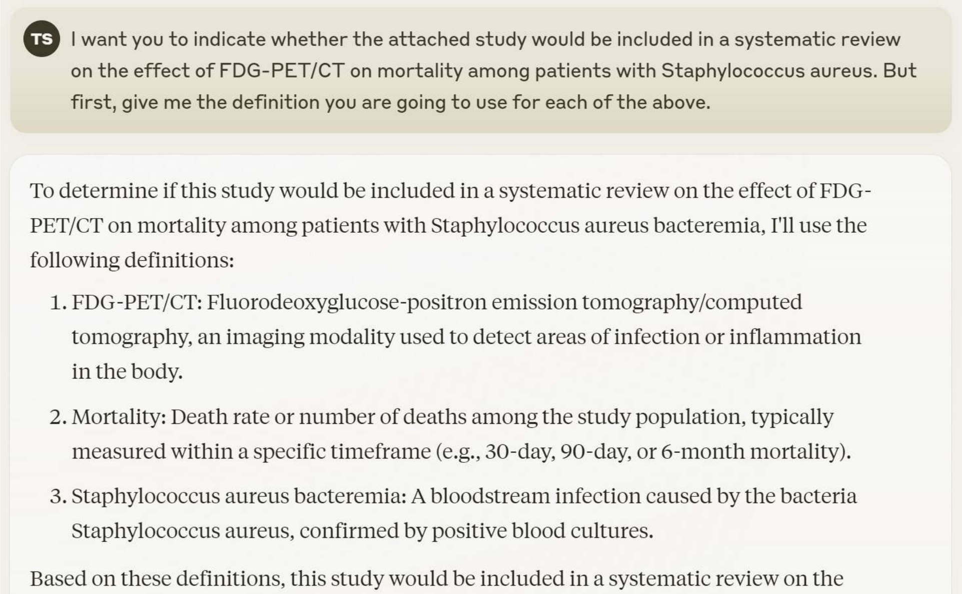
Example PICO 1: Ghanem-Zoubi et al.

- P- patients with Staphylococcus aureus
- I- Imaging with FDG-PET/CT
- C- none
- O- mortality

Example prompt: Ghanem-Zoubi et al.

I want you to indicate whether the attached study would be included in a systematic review on the effect of FDG-PET/CT on mortality among patients with Staphylococcus aureus. But first, give me the definition you are going to use for each of the above.

Claude's response



I want you to indicate whether the attached study would be included in a systematic review on the effect of FDG-PET/CT on mortality among patients with Staphylococcus aureus. But first, give me the definition you are going to use for each of the above.

To determine if this study would be included in a systematic review on the effect of FDG-PET/CT on mortality among patients with Staphylococcus aureus bacteremia, I'll use the following definitions:

1. FDG-PET/CT: Fluorodeoxyglucose-positron emission tomography/computed tomography, an imaging modality used to detect areas of infection or inflammation in the body.
2. Mortality: Death rate or number of deaths among the study population, typically measured within a specific timeframe (e.g., 30-day, 90-day, or 6-month mortality).
3. Staphylococcus aureus bacteremia: A bloodstream infection caused by the bacteria Staphylococcus aureus, confirmed by positive blood cultures.

Based on these definitions, this study would be included in a systematic review on the

Prompt adapted from Alshami et al., 2023

Human and machine working in parallel : FT Review

Example PICO 2: Ghanem-Zoubi et al.

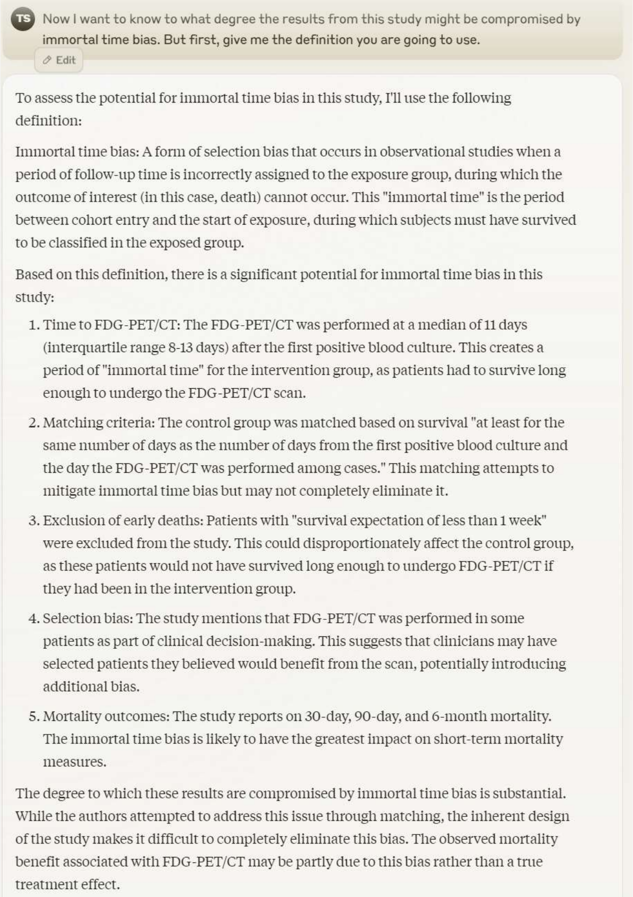
- P- patients with Staphylococcus aureus
- I- Imaging with FDG-PET/CT
- C- none
- O- mortality
- T- Observation periods across groups begun at the same time to prevent immortal time bias

Example prompt: Ghanem-Zoubi et al.

Now I want to know to what degree the results from this study might be compromised by immortal time bias. But first, give me the definition you are going to use.

Prompt adapted from Alshami et al., 2023

Claude's response:



Now I want to know to what degree the results from this study might be compromised by immortal time bias. But first, give me the definition you are going to use.

To assess the potential for immortal time bias in this study, I'll use the following definition:

Immortal time bias: A form of selection bias that occurs in observational studies when a period of follow-up time is incorrectly assigned to the exposure group, during which the outcome of interest (in this case, death) cannot occur. This "immortal time" is the period between cohort entry and the start of exposure, during which subjects must have survived to be classified in the exposed group.

Based on this definition, there is a significant potential for immortal time bias in this study:

1. Time to FDG-PET/CT: The FDG-PET/CT was performed at a median of 11 days (interquartile range 8-13 days) after the first positive blood culture. This creates a period of "immortal time" for the intervention group, as patients had to survive long enough to undergo the FDG-PET/CT scan.
2. Matching criteria: The control group was matched based on survival "at least for the same number of days as the number of days from the first positive blood culture and the day the FDG-PET/CT was performed among cases." This matching attempts to mitigate immortal time bias but may not completely eliminate it.
3. Exclusion of early deaths: Patients with "survival expectation of less than 1 week" were excluded from the study. This could disproportionately affect the control group, as these patients would not have survived long enough to undergo FDG-PET/CT if they had been in the intervention group.
4. Selection bias: The study mentions that FDG-PET/CT was performed in some patients as part of clinical decision-making. This suggests that clinicians may have selected patients they believed would benefit from the scan, potentially introducing additional bias.
5. Mortality outcomes: The study reports on 30-day, 90-day, and 6-month mortality. The immortal time bias is likely to have the greatest impact on short-term mortality measures.

The degree to which these results are compromised by immortal time bias is substantial. While the authors attempted to address this issue through matching, the inherent design of the study makes it difficult to completely eliminate this bias. The observed mortality benefit associated with FDG-PET/CT may be partly due to this bias rather than a true treatment effect.

Example 1 Discussion

Questions to consider as subgroups:

- What did you like/dislike about the AI's performance?
- What did you like/dislike about the exercise?
- Would you describe this approach as faster, slower, or about the same as conducting the steps without AI assistance?
- Would you describe this screening and data extraction with AI assistance as difficult?
- Under what conditions might you expect to get better or worse results from AI assistance?

Human and Machine working in parallel: Example 2

This time you get to choose which article to present to the AI. Please go to the internet, pull up a familiar open-access empirical study, and download the PDF version. You will share it with the AI during this next exercise.

Human and machine working in parallel : Extraction

Example 2 extraction prompt:

Please provide in a table the following information from this article. The table should have two columns, with the elements in the first column being the list of elements I provide here, and the second column being the corresponding data as found in the article. Here is the list: authors, year of publication, sample size, study design, average age, percent of sample that is female, how the presence of [variable] was measured, when the observation period began and ended, the reported effect of [treatment] on [outcome], inclusion criteria, and exclusion criteria.

Additional modifications

- Notice how Claude is extracting your data, and see if you can improve its performance by clarifying any misunderstandings.
- Other elements to extract can be added if you wish. Feel free to adapt the prompt and see how Claude performs.
- Sometimes an additional helpful command is "Add a third column to the table you just made which includes the page number and first 5 words of the paragraph or table where this information was found."

Prompt adapted from Alshami et al., 2023

Human and machine working in parallel : Screening

Example 2 screening prompt:

I want you to indicate whether the attached study would be included in a systematic review on the effect of [Intervention] compared to [Comparator] on [Outcome] among [Population]. But first, give me the definition you are going to use for each of the above.

Additional modifications

- Pay attention to how Claude is defining your PICO elements, and clarify for it any misunderstandings.
- Timing, study design, other criteria can be added if you wish. Feel free to adapt the prompt and see if Claude's response changes.
- Sometimes an additional helpful command is "Tell me what text in the manuscript you are using to make this decision. Include in your response where it is located. Include page number."

Example 2 Discussion

Questions to consider as subgroups:

- How did the AI perform on this second article?
- If it performed differently, why do you think this might be?
- What modifications did you or could you make to the prompt to help improve the AI's performance?

Global Evidence Summit

ebq
Evidence-based quality

Clinical Guidelines Resources

Overall Discussion

- How likely are you to promote the use of AI in your area?
- How central a role could AI play in the SR process?
- Under what conditions might AI performance vary?
- What additional work is needed in this area?
- What other changes might AI incorporation require?

Using evidence. Improving lives.

References

- Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making. Sociological Methods & Research, 50(2), 837-865. https://doi.org/10.1177/0049124118799372
- Beresford, M., Wutich, A., du Bray, M. V., Ruth, A., Stotts, R., Sturtz Sreetharan, C., & Brewis, A. (2022). Coding Qualitative Data at Scale: Guidance for Large Coder Teams Based on 18 Studies. International Journal of Qualitative Methods, 21. https://doi.org/10.1177/16094069221075860
- Fabiano, N., Gupta, A., Bhambra, N., Luu, B., Wong, S., Maaz, M., Fiedorowicz, J. G., Smith, A. L., & Solmi, M. (2024). How to optimize the systematic review process using AI tools. JCPP Advances, 4(2), e12234. https://doi.org/10.1002/jcv2.12234
- Hanegraaf P, Wondimu A, Mosselman JJ, et al. (2024). Inter-reviewer reliability of human literature reviewing and implications for the introduction of machine-assisted systematic reviews: a mixed-methods review BMJ Open; 14:e076912. doi: 10.1136/bmjopen-2023-076912
- Kebede MM, Le Cornet C, Fortner RT. (2023). In-depth evaluation of machine learning methods for semi-automating article screening in a systematic review of mechanistic literature. Res Syn Meth, 14(2): 156-172. doi:10.1002/jrsm.1589
- Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Syn Meth, 15(4): 616-626. doi:10.1002/jrsm.1715
- Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D. et al. (2023). Human-in-the-loop machine learning: a state of the art. Artif Intell Rev 56, 3005–3054. https://doi.org/10.1007/s10462-022-10246-w
- Pijls, B. L. (2024). Machine Learning assisted systematic reviewing in orthopaedics, Journal of Orthopaedics, 48: 103-106. https://doi.org/10.1016/j.jor.2023.11.051.
- Van Dijk, S. H. B., Brusse-Keizer M. G. J., Bucsan, C. C., van der Palen, J., Doggen, C. J.M., & Lenferink, A. (2023). Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open, 13(7): e072254.

Global Evidence Summit

ebq
Evidence-based quality

Clinical
Guidelines
Resources

Thank you!

Email: tomschofield@ebqconsulting.com
Website: https://ebqconsulting.com/

Email: cwolfkiel@indicoebm.com
Website: https://indicoebm.com/

Email: jheald@idsociety.org

Using evidence. Improving lives.

Post survey