Supplementary Material*

Table of Contents

* This supplementary material was provided by the authors to give readers further details on their article. The material was not copyedited.

Supplementary Methods

Datasets and data acquisition

We acquired the complete set of citations retrieved from electronic search from 10 distinct SRs spanning eight unique clinical domains through purposeful convenience sampling based on Oxford CEBM SR question types. To compile these datasets, we engaged with systematic review investigators at the University of Calgary and the University of Toronto. We extracted study information concerning review objectives from the published manuscript or PROSPERO protocol, and contacted study authors for internal reviewer inclusion/exclusion criteria. Study objectives and eligibility criteria for each review are reported below in Supplementary Note 1.

In brief, 'SeroTracker' (ST) was a living SR of prevalence, exploring observational cohort studies that reported single-estimate prevalence in the context of COVID-19.[1] 'Reinfection' was an SR of intervention benefits that assessed the comparative effectiveness of vaccination and past COVID-19 infection relative to past COVID-19 infection alone in observational studies reporting associations.[2] 'PA-Outcomes' was an SR of intervention benefits comparing clinical outcomes between surgery and medication treatments in patients with primary aldosteronism.[3] 'PA-Testing' was an SR of diagnostic test accuracy evaluating guideline-recommended confirmatory tests (i.e., saline infusion test, salt loading test, fludrocortisone suppression test, and captopril challenge test) relative to a reference standard.[4] 'Sepsis' was an SR of intervention benefits that assessed the comparative effectiveness and safety of fludrocortisone plus hydrocortisone, hydrocortisone alone and placebo/usual care in adults with septic shock.[5] 'Spinal' was an SR of diagnostic test accuracy that evaluated the efficacy of intraoperative neurophysiological monitoring among patients undergoing spine surgery for any indication.[6] 'Calcium-HA' was an SR of intervention benefits that compared the outcomes of routine calcium administration to no calcium administration for cardiac arrest in adults or children.[7] 'Infant-NO' was an SR of prognosis that assessed whether an immediate response to inhaled nitric oxide therapy was associated with reduced mortality in preterm infants with hypoxemic respiratory failure and pulmonary hypertension.[8] 'Meds-HA' was an SR of SRs, covering intervention benefits, that identified medications that affected hospital admissions.[9] 'SVCF' was an SR of prognosis that evaluated the association of low SVC flow, diagnosed in the first 48 hours after birth echocardiography, with neurological morbidity and mortality, among very preterm neonates.[10]

The complete set of citations retrieved from electronic search for each review was downloaded from Covidence (Veritas Health Innovation, Melbourne, Australia), a systematic review screening software. Citations were stored in csv files, containing reference information (title, authors, DOI, etc.) and abstract content. To obtain this complete set of citations, we downloaded all 'Included' (included articles after full-text screening), 'Excluded' (excluded articles during full-text screening), and 'Irrelevant' (excluded articles at abstract screening) articles from Covidence. The 'Excluded' and 'Irrelevant' citations were collated to form our excluded article dataset (complete set of citations

3

excluded during review screening). Duplicate citations, and citations missing identifier (DOI) or abstract content were discarded.

As full-text articles were not provided in Covidence, we obtained all free full-text articles with the PMC ID Converter API to convert abstract DOIs into Pubmed IDs (PMIDs), followed by the BioC API for PMC to obtain XML full-text files from each abstract PMID. Full-texts not available on PMC were excluded.

The datasets described here (complete set of citations retrieved from electronic search for each review), and all associated metadata, are available at (https://github.com/JZSang/srma). We invite other researchers to contribute to our growing *BenchSR*.

Testing Methodology

To determine the minimum sensitivity ('included') and specificity ('excluded') sample size for our evaluations during prompt development, we used the Cochran's sample size formula.[11]

For our minimum sensitivity ('included') sample size, we set our desired confidence level to 95% (p=0.05), Margin of Error (MoE) to 5%, and estimated a model sensitivity of 85%. We set our model sensitivity and specificity at 85% based on previous literature estimates.[12,13] This resulted in a minimum sensitivity sample size of 196 (included) articles.

For our minimum specificity ('excluded') sample size, we set our desired confidence level to 95% (p=0.05), Margin of Error (MoE) to 5%, and estimated a model specificity of 85%. We set our model specificity at 85% based on previous literature estimates.[12,13] This resulted in a minimum specificity sample size of 196 (excluded) articles.

We utilized data from the SeroTracker SR (n=130k excluded abstracts, n=3000 included abstracts) to randomly derive balanced sets of included and excluded abstracts for our train, and validation samples (Train: n=200 included, n=200 excluded; Validation: n=200 included, n=200 excluded), in line with our above sample size calculation. Furthermore, we randomly sampled sets of included and excluded abstracts for our few-shot GPT-CoT prompting to prevent cross contamination (GPT-CoT: n=100 included, n=100 excluded). The train and validation samples were used in our prompt development cohort.

Following prompt development, we tested the performance of our optimized prompting strategy on the ST Test sample, and 9 other SR abstract datasets to model its real-world performance and generalizability across different SR domains. These datasets were held-out from any prompt engineering steps and represent our testing cohort.

Due to the large number of citations in the ST dataset, we opted to perform a random sample of n=10,000 articles for our abstract ST test sample. Here, we set the number of 'included' articles based on the existing prevalence of 'included' articles in the complete set of citations (2.1%). This resulted in 210 'included' articles and 9790 'excluded' articles. For the 9 other SRs, we performed entire sampling, and included all 'included' and 'excluded articles. Although we only use a subset of articles in our ST test sample, our test sample of 10,000 exceeds the total number of citations retrieved in all 9 other SRs.

We replicated the same procedures for full-text evaluations. In brief, we randomly sampled the SeroTracker SR (n=4737 excluded, n=1672 included PMC-scraped full-texts) to derive our training and validation samples (Train: n=200 included, n=200 excluded; Validation: n=200 included, n=200 excluded). The SeroTracker test sample was composed of all remaining articles (Test: n=1297 included, n=4637 excluded). We did not perform few-shot experiments in our full-text evaluations due to context length limitations.

Following our prompt-optimization procedures, we tested the performance of our *ISO-ScreenPrompt* prompting strategy for full-text screening on the SeroTracker Test sample, and 9 other SR datasets. For the 9 other SRs, we performed entire sampling, and included all 'included' and 'excluded' free PMC full-texts.

Few-shot Prompting and Ablation Studies

Our few-shot examples were randomly sampled from the above-described GPT-CoT split. Unless specified, our few-shot prompts used k=10 examples (5 included, 5 excluded), in agreement with MedPrompt and general prompting guidelines.[14] We refer to this distribution of labeled examples as 'balanced' as there is an equal number of included and excluded few-shot examples.

We hypothesized that we could modulate model specificity and sensitivity by adjusting the ratio of included and excluded few-shot examples. For example, an 'inclusion-favored' distribution of examples (9 included, 1 excluded) might enhance the sensitivity of the model, and vice versa. We performed additional experiments exploring the effect of altering the distribution of our labeled few-shot examples. Here, we randomly selected 9 included and 1 excluded examples for our 'inclusion-favored' prompt; and 1 included and 9 excluded examples for our 'exclusion-favored' prompt. Results are reported in Supplement Figure 1b and Supplement Table 4.

We also perform additional ablation experiments with our prompts, to better understand the relative contribution of each prompt iteration in Abstract ScreenPrompt and ISO-ScreenPrompt. For example, in Abstract ScreenPrompt, we iteratively test the effect of adding SR objectives, or adding 'Abstract' screening instructions. For ISO-ScreenPrompt, we iteratively test prompt

repositioning experiments ('init', 'fin', 'init +fin-instructions', 'init + fin'). These results are reported in Supplement Figure 1a, 2a, and Supplement Table 3 and Supplement Table 8.

Time and Cost Analysis

Based on previous studies, the time required to screen a single abstract ranges from 20-461 seconds,[15–17] and 4.3-20 minutes for a single full-text article.[17,18] Our literature search was as follows:

Database	Search terms	Number of citations
Pubmed	("systematic review" OR "meta-analysis" OR "evidence synthesis" OR "scoping review") AND ("screening" OR "abstract screening" OR "full-text screening" OR "study selection") AND ("time" OR "time measurement" OR "duration" OR "efficiency" OR "screening time") AND ("seconds" OR "minutes")	237
Google	("systematic review" OR "meta-analysis" OR "evidence synthesis" OR "scoping review") AND ("screening" OR "abstract screening" OR "full-text screening" OR "study selection") AND ("time" OR "time measurement" OR "duration" OR "efficiency" OR "screening time") AND ("seconds" OR "minutes")	100*
*The first 100 citations were screened.		

We aligned with Perlman-Arrow et al.[16] due to our use of the same ST dataset and set the screening time at 30 seconds per abstract. We took the median time of 12 minutes 9 seconds per full-text article, based on the literature estimates retrieved from our search.

LLM API and LLM Evaluations

We used GPT3.5-Turbo-0125 (GPT3.5), GPT4-0125-preview, GPT4-Turbo-0409, GPT4o-0513, Gemini Pro, open Mixtral-8x22-0424, Mistral-Large-0224, and Claude-3.5-Sonnet-1022 models and compared their performance for abstract and full-text screening with our optimized 'Abstract ScreenPrompt' and 'ISO-ScreenPrompt prompts', respectively. For all models, we set the maximum tokens to 2048, and used the advised default settings. GPT model settings were set to temperature=1, top_p=1, frequency_penalty=0, presence_penalty=0. The Gemini pro model settings were set to temperature=1, top_p=1. The Mixtral-8x22-0424, and Mistral-Large-0224 model settings were set to temperature=0.7, top_p=1. Claude model settings were set to temperature=1, top_p=1, top_k=250. We note that for Gemini Pro, the default temperature was changed to temperature=0.9 as of May 2024 (after our testing). We set the seed for all of our LLM models at 0.

The temperature and top_p ranges for all of the models used in our investigation ranges from 0-2, with 0 being most deterministic and 2 being the most random and diverse. We chose to keep temperature and top_p parameters at 1 for two reasons. Firstly, previous research has indicated that model accuracy for problem-solving tasks does not significantly differ within the temperature range of 0 to 1.[19] Second, because we conducted a cross-model evaluation (i.e., comparing Gemini, GPT, Mistral models), we used default temperature settings to avoid unnecessary variation. For instance, the default temperature for Gemini is set to 0.9, while the default temperature for GPT is set to 1. It is unclear whether setting the same temperature value across different models results in the same degree of stochasticity due to potential differences in their underlying architectures and training data. Therefore, using the default settings for each model ensured consistency and comparability in our evaluations

To evaluate the model responses, we required an output containing an 'evaluation token': either "XXX" (exclude) or "YYY" (include). When responses contained both "XXX" and "YYY", we checked the last 500 characters of the output and used the final instance of the token.

We conducted our initial runs synchronously, sending requests directly to the API and waiting for responses in the same network call. This approach allowed us to easily retry individual calls to a model's API. For each article, if a response was interrupted by a technical error (rate limit, disconnect, timeout) or was missing an 'evaluation token', we retried the instance up to three times. If all three attempts failed, we deemed the request as 'unparseable' and disqualified the request from the total count, occasionally reducing the set by 1-2 articles. During the writing of this paper, OpenAI released the Batch API, which significantly reduced costs and improved run times. Shifting to this infrastructure meant we could no longer efficiently retry individual abstracts, but the new system greatly reduced technical errors and rarely resulted in missing "evaluation tokens".

DATA AVAILABILITY
Researchers can access our data via the following github repository (https://github.com/JZSang/srma).

CODE AVAILABILITY
All code used for experiments in this study can be found in a github repository (https://github.com/JZSang/srma).

Supplementary Figures

Supplement Figure 1: Infographic of proposed SR screening workflows: (i) LLMs as a single reviewer and (ii) LLMs as a 'pre-screen' to dual human screening.
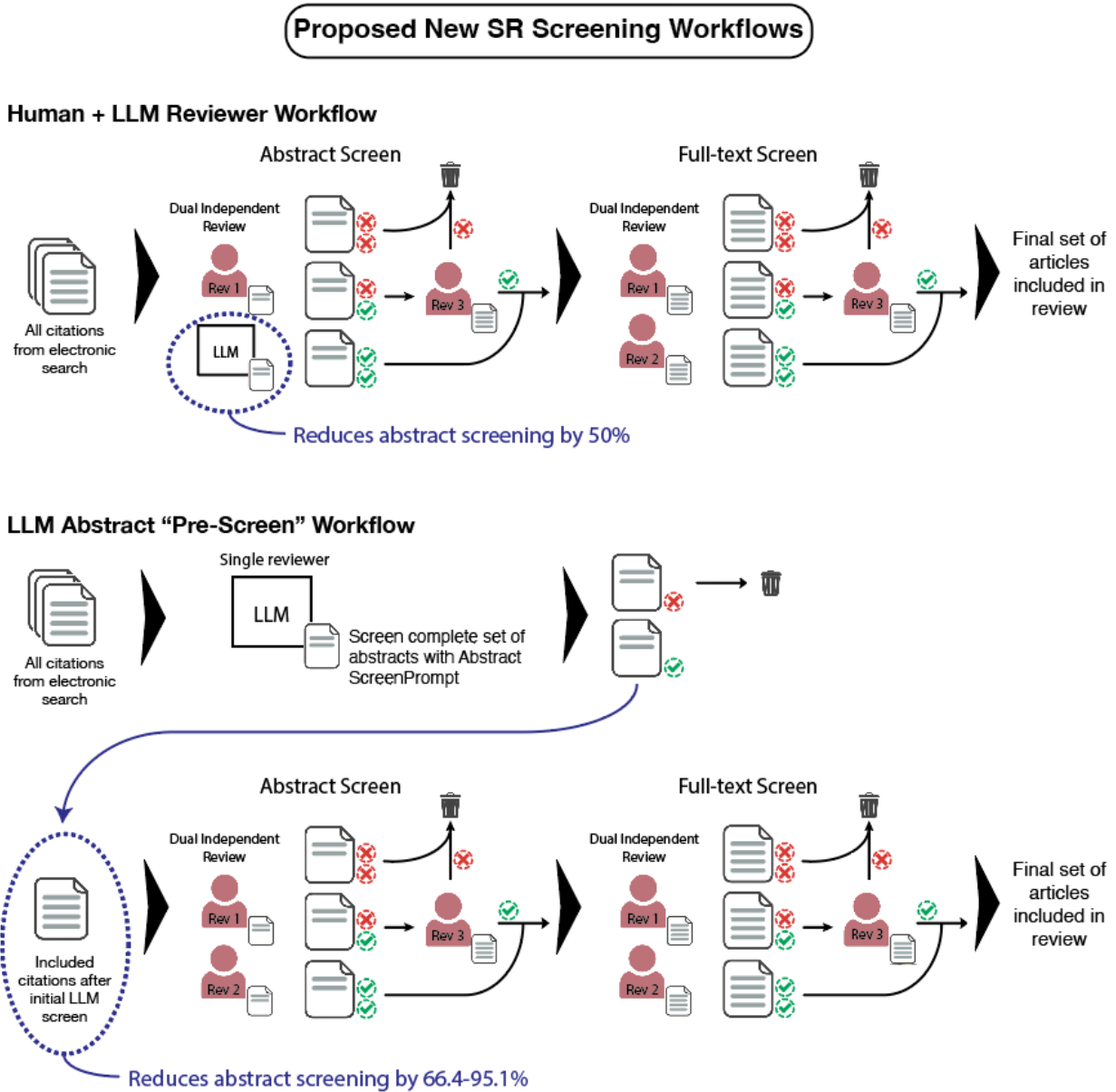
Supplement Figure 2: Abstract screening prompt optimization.

a) Performance comparison of different abstract prompting methodologies relating to the addition of study objectives and abstract-specific considerations on the SeroTracker (ST) training dataset, (n=400) showing accuracy and sensitivity. Error bars represent 95% CIs for binomial proportions. b) Performance comparisons of different few-shot prompting methodologies on the ST training dataset, showing accuracy and sensitivity. Prompts compare differing proportions of inclusion-labeled and exclusion-labeled few-shot examples. Error bars represent 95% CIs for binomial proportions.
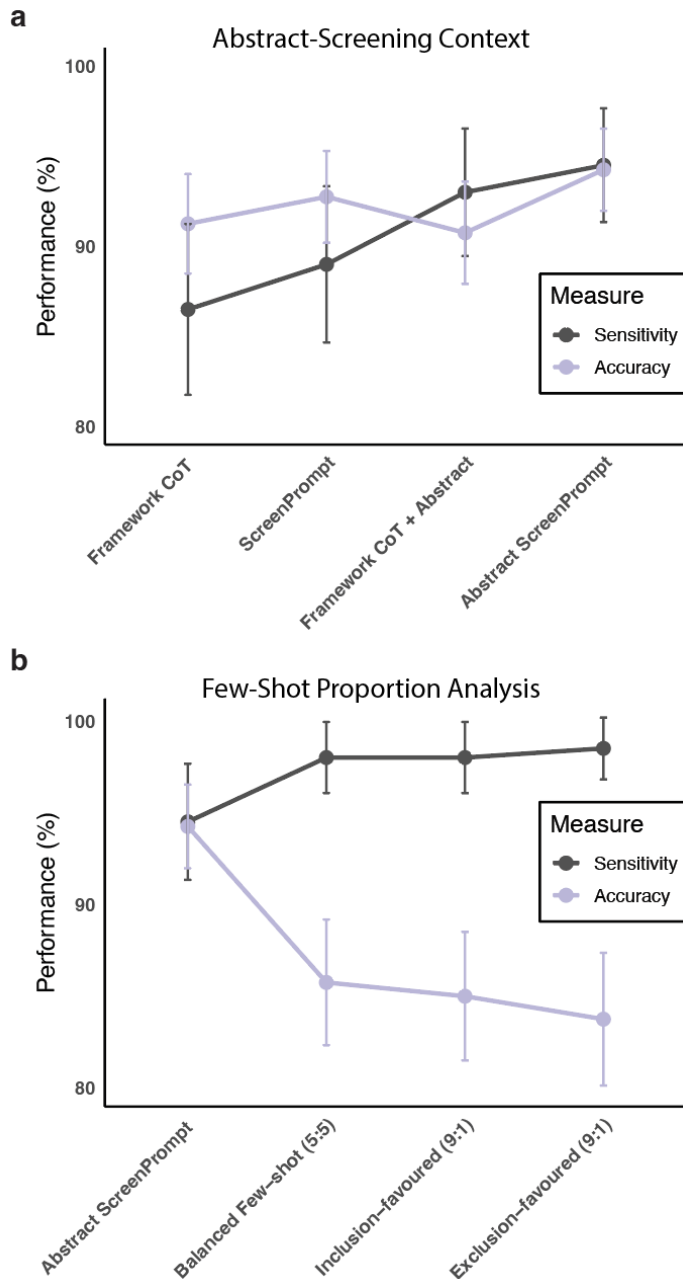
Supplement Figure 3: Full-text screening prompt optimization.
a) Performance comparison of different full-text prompting methodologies with modifications in prompt structure on the ST training dataset (n=400), showing accuracy and sensitivity. Error bars represent 95% CIs for binomial proportions. b) Performance comparison of abstract screening performance with modifications in prompt structure on the ST training dataset, showing accuracy and sensitivity. Error bars represent 95% CIs for binomial proportions.
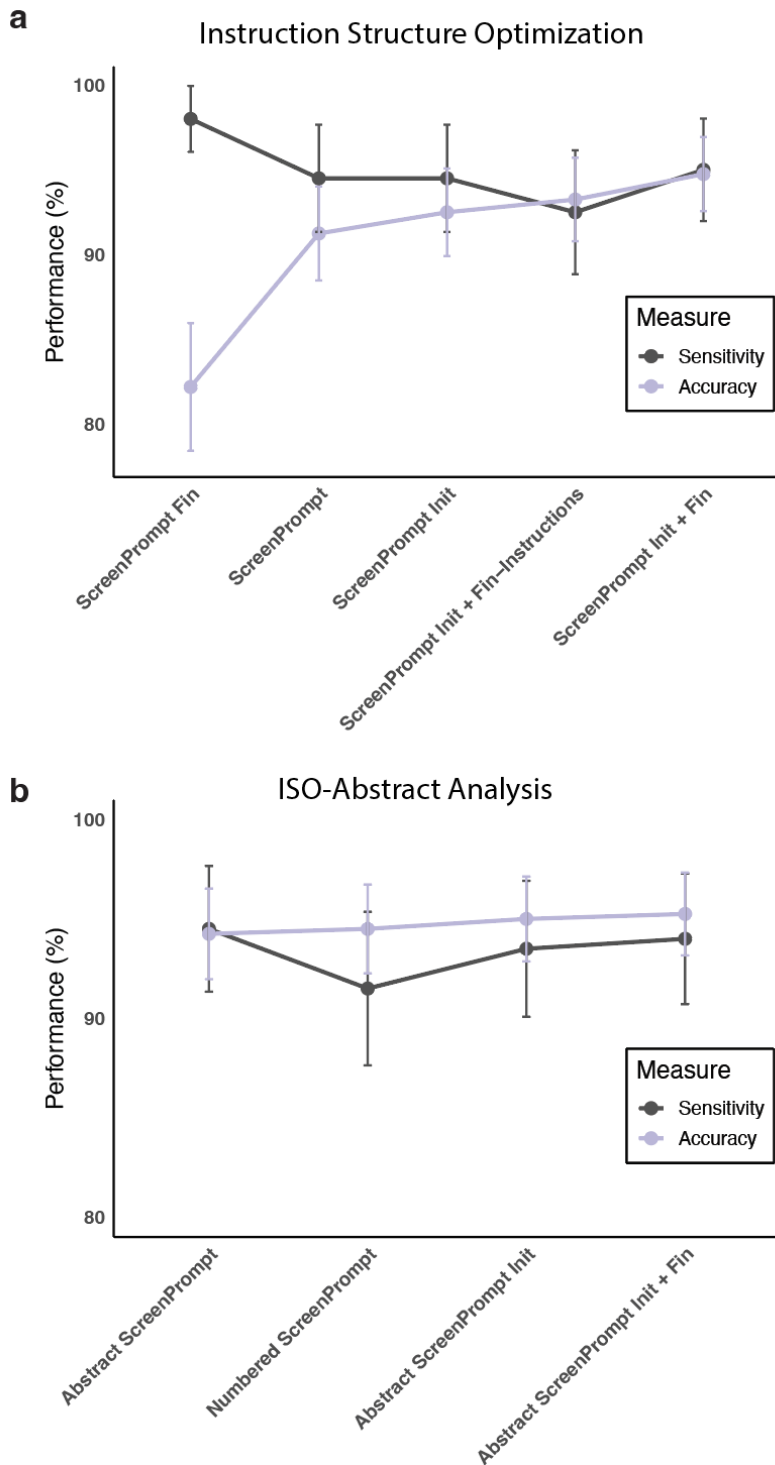
Supplementary Tables

Supplement Table 1: Abstract and Full-text Screening Prompts

Prompting Technique and Definitions	Prompt Template.
	Prompt modules are enclosed in {} and were not inserted in the prompt.
	Italics reflect prompt module content.
	Bolded sections highlight iterative prompt changes.
Zero-shot *Adapted from Guo et al.[12] Models are prompted with only instructions for completing the task at hand, including the required context directly related to the task, i.e. abstract text to analyze. No additional context or examples are provided. Our zero-shot prompt was adapted from Guo et al.[12] and we similarly instructed the LLM to return a single token output with its final decision.	{Pre-Prompt} *You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper. Use the criteria below to inform your decision. If any exclusion criteria are met or not all inclusion criteria are met, exclude the article. If all inclusion criteria are met, include the article.* {Inclusion Criteria} {Exclusion Criteria} {Abstract in investigation} {Instructions} *Only type "YYY" for included articles or "XXX" for excluded articles to indicate your decision. Do not type anything else.* {Model output}
Random few-shot (k=10)	{Pre-Prompt}

Models are prompted with the instruction-task at hand, with the addition of (k=n) labeled examples relevant to the task at hand. For our purposes, the examples were randomly selected and accurately labeled as included ('YYY') or excluded ('XXX'). We set k=10 (5 included, 5 excluded), in agreement with MedPrompt and general prompting guidelines.[14]	*You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper. Use the criteria below to inform your decision. If any exclusion criteria are met or not all inclusion criteria are met, exclude the article. If all inclusion criteria are met, include the article.* {Inclusion Criteria} {Exclusion Criteria} **{Example Include n= 5}** **{Example Exclude n= 5}** {Abstract in investigation} {Instructions} *Only type "YYY" for included articles or "XXX" for excluded articles to indicate your decision. Do not type anything else.* {Model output}
Zero-shot CoT Models are prompted with the instruction-task at hand, along with additional natural language statements, such as "*Let's think step by step*" to encourage the	{Pre-Prompt} *You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper. Use the criteria below to inform your decision. If any exclusion criteria are met or not all inclusion criteria are met, exclude the article. If all inclusion criteria are met, include the article.* {Inclusion criteria}

model to generate intermediate reasoning steps before generating a final answer.[20]	{Exclusion Criteria} {Abstract in investigation} {Instructions} ***Let's think step by step for why an article should be included or excluded.*** *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
Zero-shot Framework CoT We devised a new prompting approach, termed 'Framework CoT,' wherein we deliberately prompt the model to reason against each criterion. Similar to zero-shot prompting, no additional context or examples are provided.	{Pre-prompt} *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {Inclusion criteria} {Exclusion Criteria} {Abstract in investigation} {Instructions} *# Instructions*

	We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met. *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
ScreenPrompt We included additional well-defined study objectives (verbatim from published manuscripts) to our Framework CoT prompt to better orient our prompt to screening tasks.	{Pre-prompt, including **objectives**} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {Inclusion criteria} {Exclusion Criteria} {Abstract in investigation}

	{Instructions, including task considerations}
	# Instructions *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
Abstract ScreenPrompt We further incorporated additional context that acknowledged the inherent content limitations of abstracts and goals of inclusivity to our ScreenPrompt prompt to better orient our prompt to the task of abstract screening.	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {Inclusion criteria} {Exclusion Criteria}

	{Abstract in investigation} {Instructions, including abstract considerations} *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* ***Studies that may not fully align with the primary focus of our inclusion criteria but provide data or insights potentially relevant to our review deserve thoughtful consideration. Given the nature of abstracts as concise summaries of comprehensive research, some degree of interpretation is necessary.*** ***Our aim should be to inclusively screen abstracts, ensuring broad coverage of pertinent studies while filtering out those that are clearly irrelevant.*** *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
Few-shot GPT-CoT Abstract ScreenPrompt (k=10) We prompted our model with *Abstract ScreenPrompt*, but also	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2*

included additional examples that contained *Abstract ScreenPrompt* GPT-generated reasoning. We discarded answers that did not match the ground truth label to uphold 'correctness' in example reasoning. This approach was adapted from MedPrompt, which has suggested that GPT-generated CoT reasoning can outperform human experts, as well as automate the CoT example process. We set k=10 (5 included, 5 excluded), in agreement with MedPrompt and general prompting guidelines.[14]

infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

{Inclusion criteria}

{Exclusion Criteria}

{GPT-CoT Include n= 5}

{GPT-CoT Exclude n= 5}

{Abstract in investigation}

{Instructions, including abstract considerations}
 # Instructions
We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.

Studies that may not fully align with the primary focus of our inclusion criteria but provide data or insights potentially relevant to our review deserve thoughtful consideration. Given the nature of

	abstracts as concise summaries of comprehensive research, some degree of interpretation is necessary. *Our aim should be to inclusively screen abstracts, ensuring broad coverage of pertinent studies while filtering out those that are clearly irrelevant. We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
Full-text	
Numbered ScreenPrompt: We preserve the semantic content of our prompt and add number symbols for each individual inclusion and exclusion sub-criterion. We removed meta-criteria headings (i.e., population, intervention, etc.) where applicable.	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {**Numbered** Inclusion criteria} *1. Humans of any age* …

	11. *Reports the locations at which the study took places such that they could be categorized as neighbourhood, city, state/province/territory, or country* {**Numbered** Exclusion Criteria} ***1.*** *Non-human (e.g., in silico, animal, in vitro)* … ***10.*** *Does not report the location at which the study took place* {Full-text in investigation} {Instructions} *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
ScreenPrompt (Init): We modify our 'prompt structure' by appending all of our prompt elements ({Objectives},	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2*

{Inclusion criteria}, {Exclusion criteria}, {Instructions}) to the start of the prompt, before the {Full-text} content. (Supplement Table 8)	*infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {Inclusion criteria} {Exclusion Criteria} **{Instructions}** *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Full-text in investigation} {Model output}
ScreenPrompt (Fin):	{Full-text in investigation}

We modify our 'prompt structure' by appending all of our prompt elements to the end of the prompt, after the {Full-text} content. (Supplement Table 8)	**{Pre-prompt, including objectives}** *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* **{Inclusion criteria}** **{Exclusion Criteria}** **{Instructions}** *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* **{Model output}**

ScreenPrompt (Init + Fin-Instructions): We modify our 'prompt structure' by appending all of our prompt elements to the start of the prompt, before the {Full-text} content. We additionally append {Instructions} after the {Full-text} content. (Supplement Table 8)	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {Inclusion criteria} {Exclusion Criteria} **{Instructions}** *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Full-text in investigation}

	{Instructions} *# Instructions* *We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.* *We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.* {Model output}
ScreenPrompt (Init + Fin): We modify our 'prompt structure' by appending all of our prompt elements to the start of the prompt (before the {Full-text} content) and end of the prompt (after the {Full-text} content).	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {Inclusion criteria} {Exclusion Criteria}

{Instructions}

 # Instructions

We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.

We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.

{Full-text in investigation}

{Pre-prompt, including objectives}

Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

{Inclusion criteria}

{Exclusion Criteria}

	{Instructions}
	# Instructions
	We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.
	We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.
ISO-ScreenPrompt: We apply a combination of our Numbered ScreenPrompt where each individual inclusion/exclusion sub-criterion is numbered, and our (Init + Fin) prompt structure, where all prompt elements are appended to the start and ends of the prompt.	{Pre-prompt, including objectives} *Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.* *The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:* {**Numbered** Inclusion criteria} {**Numbered** Exclusion Criteria} {Instructions}

 # Instructions

We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.

We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.

{Full-text in investigation}

{Pre-prompt, including objectives}

Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

{Numbered Inclusion criteria}

{Numbered Exclusion Criteria}

{Instructions}

Instructions

We now assess whether the paper should be included from the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.

We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.

{Model output}

Supplement Table 2: Abstract prompt engineering performance

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	Zero-shot -	65.0% 260/400	30.0% 60/200	**100.0%** **200/200**
	Random Few-shot Balanced: 5 include 5 exclude	78.0% 312/400	56.0% 112/200	100.0% 200/200
	Zero-shot Chain of Thought (CoT)	86.3% 345/400	74.5% 149/200	98.0% 196/200
	Zero-shot framework CoT	91.3% 365/400	86.5% 173/200	96.0% 192/200
	ScreenPrompt	92.8% 371/400	89.0% 178/200	96.5% 193/200
	Abstract ScreenPrompt	**94.3%** **377/400**	94.5% 189/200	94.0% 188/200
	Few-shot GPT-CoT ScreenPrompt Balanced: 5 include 5 exclude	85.8% 343/400	**98.0%** **196/200**	73.5% 147/200
Validation (n=200 include, n=200 exclude)	Abstract ScreenPrompt	94.3% 377/400	96.0% 192/200	92.5% 185/200

28

Supplement Table 3: Exemplar Wrong Outputs during Abstract Prompt Engineering

Prompt	Example LLM Output	Author Feedback
Zero-shot CoT:	{Model output} **Population**: 1. The study involves humans of any age. [Meets criteria] 2. The subjects include those with suspected disease without requiring them to have been deliberately sampled for being antigen positive. [Meets criteria] **Study Design**: 1. This is a cross-sectional serologic survey. It fits the inclusion criteria for study design as it is one of the specified designs interested in measuring seroprevalence. [Meets criteria] 2. Serology measurements were presumably taken at a single time point as common in cross-sectional designs. [Meets criteria] **Special Design**: 1. This paper is not a systematic review or meta-analysis but rather a primary seroprevalence study. However, non-systematic reviews/meta-analyses are not listed as exempt from consideration; the inclusion criteria express a specific interest in systematic reviews and meta-analyses without excluding primary studies. [Meets criteria] **Sampling**: 1. The method of sampling is not entirely detailed in the abstract, but given that it's a cross-sectional study among a defined number of participants (888), it suggests a deliberate approach to sampling, which is compliant with the inclusive approach to any sampling method. [Meets criteria] **Types of Evidence**: 1. Assuming this is a published academic article (given its detailed results and conclusions, along with the referencing of copyright by Wiley & Sons Ltd.), it meets the criteria for the types of evidence being sought. [Meets criteria] **Outcome Measures**: 1. The study reports a seroprevalence estimate, with 30% of the study volunteers having detected neutralizing antibodies. [Meets criteria] 2. The number of participants enrolled in the study is clearly stated as 888. [Meets criteria] 3. While the study sampling end date/week is not explicitly mentioned, this detail might be present in the full text; however, for abstract screening, its absence is noted but its potential availability in the full report could still mean the study might meet the criteria upon full review. [Unclear] 4. The study clearly reports the location as being in Egypt, fitting the requirement to categorize the location of the study. [Meets criteria] Given that all but one of the clear inclusion criteria are met with only the study sampling end date/week not explicitly stated in the abstract, there's a stronger leaning towards inclusion, with the assumption that further details necessary to fully satisfy all criteria might be found in the full text. **Conclusion: YYY**	Output did not assess/evaluate exclusion criteria → Led to Framework CoT
Zero-Shot Framework CoT	Let's evaluate the abstract provided against the inclusion and exclusion criteria systematically. ### Inclusion Criteria Evaluation: 1. **Population:** The study includes humans of any age, which meets the criterion. It doesn't specify excluding only those with suspected or confirmed	Erroneously inferred study objectives → Led to addition of

COVID-19, suggesting broader inclusion, thus meeting this requirement. 2. **Study design:** The study design focuses on the qualitative assay of immunoglobulins G and M in diagnosed patients, aiming to assess transmissibility and the possibility of reinfection. This can be interpreted as part of a broader sero-survey effort, but it appears more focused on specific diagnostic outcomes rather than prevalence estimation. It lacks a clear mention of being a cross-sectional, repeated cross-sectional, or cohort study as defined by our criteria. 3. **Special design:** The study does not present itself as a systematic review or meta-analysis of seroprevalence studies. 4. **Sampling:** The method of sampling is not explicitly described, but the study included a total of 553 patients, indicating a form of sampling was done. 5. **Types of evidence:** Presumably, this is published or unpublished academic literature, considering it has authorship and an abstract, thus meeting this criterion. 6. **Outcome measures:** The abstract reports on IgG and IgM levels in patients, implying an assessment of seroprevalence to some extent. However, it does not clearly state it reports a seroprevalence estimate in terms of the proportion of the population with detectable antibodies. The number of participants is reported, but it lacks clear information on study sampling end date/week or detailed location categorization. ### Exclusion Criteria Evaluation: 1. **Population:** This study is conducted in humans, thus not meeting any of the population-based exclusion criteria directly. 2. **Study design:** The study could arguably fall under the exclusion due to its focus on evaluating immunological responses rather than explicitly being a cross-sectional or cohort design aimed at seroprevalence estimation. 3. **Types of evidence:** There's no mention of using excluded multimedia sources or unassociated dashboards. 4. **Outcome measures:** The abstract mainly discusses the detection of IgG and IgM antibodies but does not only report on antigen prevalence. However, it does not fully comply with reporting requirements around the study sampling end date/week or location detail. ### Conclusion: <u>**Given the specific focus on evaluating immunological responses in patients rather than estimating the seroprevalence across a population,**</u> and the absence of clear compliance with several crucial inclusion criteria (study design specifics, complete outcome measure reporting), the study appears to lean more towards exclusion based on its design and reported outcomes. However, considering the nature of abstract reviews, the decision leans on available information which might not fully capture the study's relevance to the systematic review aims. Given the potential relevance of immunological response data to the broader topic of seroprevalence and immunity estimation, there is a fine line of interpretation. Considering the primary focus on diagnostic outcomes as opposed to population-level antibody prevalence estimation and the critical unfulfilled aspects of both inclusion and exclusion criteria, the conclusion leans towards: XXX	SR objectives (ScreenPrompt)

Supplement Table 4: Abstract ScreenPrompt Ablation Prompting

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	Zero-shot framework CoT	91.3% 365/400	86.5% 173/200	96.0% 192/200
	Zero-shot framework CoT ++ Study Objectives (ScreenPrompt)	92.8% 371/400	89.0% 178/200	**96.5%** **193/200**
	Zero-shot framework CoT ++ Abstract Consideration	90.8% 363/400	93.0% 186/200	88.5% 177/200
	Zero-shot framework CoT ++ Abstract Consideration ++ Study Objectives (Abstract ScreenPrompt)	**94.3%** **377/400**	**94.5%** **189/200**	94.0% 188/200

31

Supplement Table 5: Few-shot GPT-CoT Label Proportion Analysis

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	Few-shot GPT-CoT ScreenPrompt Balanced: 5 include 5 exclude	**85.8%** **343/400**	98.0% 196/200	**73.5%** **147/200**
	Few-shot GPT-CoT ScreenPrompt Inclusion-favored: 9 include 1 exclude	85.0% 340/400	98.0% 196/200	72.0% 144/200
	Few-shot GPT-CoT ScreenPrompt Exclusion-favored: 1 include 9 exclude	83.8% 335/400	**98.5%** **197/200**	69.0% 138/200

Supplement Table 6: Comparative analysis of Abstract ScreenPrompt across LLM models

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity	Cost (CAD)
ST Training n=200 include n=200 exclude	GPT4-0125-preview Abstract ScreenPrompt	**94.3%** **377/400**	**94.5%** **189/200**	94.0% 188/200	$12.54
	GPT4-Turbo-0409 Abstract ScreenPrompt	89.8% 359/400	83.5% 167/200	96.0% 192/200	$12.16
	GPT3.5 Abstract ScreenPrompt	66.7% 266/399	90.5% 180/199	43.0% 86/200	$0.37
	Gemini Pro Abstract ScreenPrompt	76.2% 301/395	67.5% 133/197	84.8% 168/198	**Free**
	Mistral-Large Abstract ScreenPrompt	89.3% 357/400	93.5% 187/200	85.0% 170/200	$2.65
	Mixtral-8x22 Abstract ScreenPrompt	84.5% 337/399	94.0% 188/200	74.9% 149/199	$4.15
	GPT4o-0513 Abstract ScreenPrompt	89.0% 356/400	79.5% 159/200	**98.5%** **197/200**	$7.37
	Claude-3.5-Sonnet Abstract ScreenPrompt	93.3% 373/400	88.0% 176/200	98.5% 197/200	$6.12

Supplement Table 7: Zero-shot vs. Abstract ScreenPrompt across LLM Models

Dataset	Model	Prompt	Accuracy	Sensitivity	Specificity
Training n=200 include n=200 exclude	GPT4-0125-preview	Abstract ScreenPrompt	**94.3%** **377/400**	**94.5%** **189/200**	94.0% 188/200
		Zero-shot	65.0% 260/400	30.0% 60/200	**100.0%** **200/200**
	GPT4-Turbo-0409	Abstract ScreenPrompt	**89.8%** **359/400**	**83.5%** **167/200**	96.0% 192/200
		Zero-shot	56.8% 227/400	13.5% 27/200	**100.0%** **200/200**
	GPT3.5	Abstract ScreenPrompt	**66.7%** **266/399**	90.5% 180/199	**43.0%** **86/200**
		Zero-shot	64.5% 258/400	**97.0%** **193/200**	32.5% 65/200
	Claude 3.5 Sonnet	Abstract ScreenPrompt	**93.3%** **373/400**	**88.0%** **176/200**	**98.5%** **197/200**
		Zero-shot	75.8% 303/400	51.5% 103/200	100% 200/200

Supplement Table 8: Full-text prompt engineering performance

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	Zero-shot	68.9% 275/399	37.7% 75/199	100.0% 200/200
	Abstract ScreenPrompt	85.5% 342/400	**98.0%** **196/200**	73.0% 146/200
	ScreenPrompt	91.3% 365/400	94.5% 189/200	88.0% 176/200
	ScreenPrompt Init + Fin	94.8% 379/400	95.0% 190/200	94.5% 189/200
	Numbered ScreenPrompt	95.3% 381/400	96.0% 192/200	94.5% 189/200
	ISO-ScreenPrompt	**95.5%** **382/400**	93.5% 187/200	**97.5%** **195/200**
Validation (n=200 include, n=200 exclude)	ISO-ScreenPrompt	96.3% 385/400	97.5% 195/200	95% 190/200

Supplement Table 9: Full-text Prompt Structure Testing

Dataset	Prompting Strategy (Structure)	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	ScreenPrompt	91.3% 365/400	94.5% 189/200	88.0% 176/200
	ScreenPrompt Init	92.5% 370/400	94.5% 189/200	90.5% 181/200
	ScreenPrompt Fin	82.2% 328/399	**98.0%** **196/200**	66.3% 132/199
	ScreenPrompt Init + Fin-Instructions	93.3% 373/400	92.5% 185/200	94.0% 188/200
	ScreenPrompt Init + Fin	**94.8%** **379/400**	95.0% 190/200	**94.5%** **189/200**

Supplement Table 10: ISO-Prompting in Abstract Screening

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	Abstract ScreenPrompt (Regular structure)	94.3% 377/400	94.5% 189/200	94.0% 188/200
	Numbered Abstract ScreenPrompt	94.5% 378/400	91.5% 183/200	97.5% 195/200
	Abstract ScreenPrompt Init	95.0% 380/400	93.5% 187/200	96.5% 193/200
	Abstract ScreenPrompt Init + Fin	95.3% 381/400	94.0% 188/200	96.5% 193/200

Supplement Table 11: Comparative analysis of ISO-ScreenPrompt across LLM models

Dataset	Prompting Strategy	Accuracy	Sensitivity	Specificity	Cost (CAD)
Training (n=200 include, n=200 exclude)	GPT4-0125-preview ISO-ScreenPrompt	95.5% 382/400	**93.5%** **187/200**	97.5% 195/200	$51.90
	GPT4-Turbo-0409 ISO-ScreenPrompt	**95.8%** **383/400**	93.0% 186/200	**98.5%** **197/200**	$51.06
	GPT3.5 ISO-ScreenPrompt	76.1% 251/330	74.1% 129/174	78.2% 122/156	$1.76
	Gemini Pro ISO-ScreenPrompt	67.3% 255/379	69.1% 134/194	65.4% 121/185	**Free**
	Mistral-Large ISO-ScreenPrompt	86.9% 345/397	77.8% 154/198	96% 191/199	$19.77
	Mixtral-8x22 ISO-ScreenPrompt	93.7% 370/395	91.9% 182/198	95.4% 188/197	$9.93
	GPT4o-0513 ISO-ScreenPrompt	95.3% 381/400	93.0% 186/200	97.5% 195/200	$27.12
	Claude-3.5-Sonnet ISO-ScreenPrompt	96.3% 385/400	93.5% 187/200	99% 198/200	$18.65

Supplement Table 12: Zero-shot vs. ISO-ScreenPrompt across LLM Models

Dataset	Model	Prompt	Accuracy	Sensitivity	Specificity
Training (n=200 include, n=200 exclude)	GPT4-0125-preview	**ISO-ScreenPrompt**	**95.5%** **382/400**	**93.5%** **187/200**	97.5% 195/200
		Zero-shot	68.9% 275/399	37.7% 75/199	**100.0%** **200/200**
	GPT4-Turbo-0409	**ISO-ScreenPrompt**	**95.8%** **383/400**	**93.0%** **186/200**	98.5% 197/200
		Zero-shot	56.8% 227/400	13.5% 27/200	**100.0%** **200/200**
	GPT3.5	**ISO-ScreenPrompt**	**76.1%** **251/330**	74.1% 129/174	**78.2%** **122/156**
		Zero-shot	72.1% 253/351	**92.8%** **168/181**	50.0% 85/170
	Claude 3.5 Sonnet	**ISO-ScreenPrompt**	**96.3%** **385/400**	**93.5%** **187/200**	**99%** **198/200**
		Zero-shot	72% 385/400	50% 100/200	94% 188/200

Supplement Table 13: Estimated Time and Cost Savings Analysis for Abstract ScreenPrompt compared to Single Human-Reviewer Screening

Dataset	Number of Abstracts	Estimated time - Human (hours)	Estimated single reviewer cost (USD)	Abstract ScreenPrompt GPT4-0125-preview cost with Batch API (USD)
SeroTracker	10000	83.33	1666.67	157.02
Reinfection	6178	51.48	1029.67	113.96
PA-Testing	7757	64.64	1292.83	130.6
PA-Outcome	4309	35.91	718.17	64.52
Meds-HA	9628	80.23	1604.67	141.81
Sepsis	4017	33.48	669.50	47.32
Spinal	1914	15.95	319.00	26.99
Infant-NO	1169	9.74	194.83	16.74
Calcium-HA	1502	12.52	250.33	18.68
SVCF	1951	16.26	325.17	17.12

Supplement Table 14: Estimated Time and Cost Savings Analysis for ISO-ScreenPrompt compared to Single Human-Reviewer Screening

Dataset	Number of PMC Full-texts	Estimated time - Human (hours)	Estimated single reviewer cost (USD)	Abstract ScreenPrompt GPT4-0125-preview cost with Batch API (USD)
SeroTracker	6409	1297.82	25,956.40	622.12
Reinfection	2752	557.28	11,145.60	351.97
PA-Testing	719	145.60	2,911.95	82.48
PA-Outcome	482	97.60	1,952.10	53.85
Meds-HA	1078	218.30	4,365.90	132.45
Sepsis	404	81.81	1,636.20	38.26
Spinal	244	49.41	988.20	28.65
Infant-NO	197	39.89	797.85	21.90
Calcium-HA	238	48.20	963.90	22.54
SVCF	167	33.82	676.35	14.53

Supplement Table 15: Estimated Time and Cost Savings with Abstract ScreenPrompt Pre-Screening

Dataset	Number of Abstracts	Citations removed with LLM pre-screen (%)	Estimated time saved (hours)	Estimated dual reviewer cost-savings (USD)	Abstract ScreenPrompt GPT4-0125-preview cost with Batch API (USD)
SeroTracker	10000	8982 (89.8%)	149.70	2994.00	157.02
Reinfection	6178	4100 (66.4%)	68.33	1366.67	113.96
PA-Testing	7757	6991 (90.1%)	116.52	2330.33	130.6
PA-Outcome	4309	3339 (77.5%)	55.65	1113.00	64.52
Meds-HA	9628	8097 (84.1%)	134.95	2699.00	141.81
Sepsis	4017	3475 (86.5%)	57.92	1158.33	47.32
Spinal	1914	1405 (73.4%)	23.42	468.33	26.99
Infant-NO	1169	1029 (88.0%)	17.15	343.00	16.74
Calcium-HA	1502	1428 (95.1%)	23.80	476.00	18.68
SVCF	1951	1777 (91.1%)	29.62	592.33	17.12

Supplementary Note 1: SR Objectives and Eligibility Criteria

The documents below represent the identical study objectives and eligibility criteria that were inserted in our prompts ('simplified objectives and inclusion/exclusion criteria'). We also report the 'Original Author Protocol' documents that were provided by the original study authors for each SR.

SeroTracker (ST)

Simplified Objectives + Inclusion/Exclusion Criteria

Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion Criteria (all must be fulfilled):

Population

1. Humans of any age

2. Including COVID-19 antigen positive persons and those with suspected disease if not deliberately sampled.

Study design

1. Sero-surveys – defined as the collection and testing of serum (or proxy such as oral fluid) specimens from a sample of a defined population over a specified period of time to estimate the prevalence of antibodies against SARS-CoV-2 as an indicator of immunity

2. Cross-sectional, repeated cross sectional, and cohort study designs, with serology measurements at single time points or repeated at multiple time points

Special design

1. Include systematic reviews and meta-analysis of seroprevalence studies for the purpose of tracking evidence synthesis efforts

Sampling

1. Any sampling method

Types of evidence

1. Published or unpublished academic literature, grey literature (government or institutional reports), or media reports. Slide deck presentations were included if we could identify the person giving the presentation and the date of the presentation

Outcome measures

1. Reports a seroprevalence estimate (proportion of the population with detectable antibodies)

2. Reports the number of participants enrolled in the study (denominator)

3. Reports study sampling end date/week

4. Reports the locations at which the study took places such that they could be categorized as neighbourhood, city, state/province/territory, or country

Exclusion Criteria (if any met then exclude):

Population

1. Non-human (e.g., in silico, animal, in vitro)

2. The study only included individuals with suspected, active, or previously diagnosed with COVID-19 using PCR, antigen testing, clinical assessment, or self-assessment

3. The study only included individuals vaccinated against SARS-CoV2

Study design

1. Study designs other than cross-sectional or cohort design: case reports, case-control studies, evaluations of serological tests, study protocols

Types of evidence

1. Multimedia sources of data (audio clips, video clips) were excluded due to the feasibility of extracting. Slide deck presentations were excluded if we could not identify the person giving the presentation and the date of the presentation

2. Dashboards not associated with a defined serology study

Outcome measures

1. Only reports incidence or prevalence of SARS-CoV-2 antigen (as opposed to antibody)

2. Does not report study sampling end date/week

3. Does not report the number of participants included in the study (sample denominator)

4. Does not report the location at which the study took place

Original author protocol

SeroTracker Inclusion/Exclusion Criteria

Criteria for including evidence (must meet all the criteria to be included)

Characteristics	Criteria for inclusion
Population	· Humans of any age o Including COVID-19 antigen positive persons and those with suspected disease if not deliberately sampled.
Study design	· Sero-surveys – defined as the collection and testing of serum (or proxy such as oral fluid) specimens from a sample of a defined population over a specified period of time to estimate the prevalence of antibodies against SARS-CoV-2 as an indicator of immunity · Cross-sectional, repeated cross sectional, and cohort study designs, with serology measurements at single time points or repeated at multiple time points
*Special design	· Include systematic reviews and meta-analysis of seroprevalence studies for the purpose of tracking evidence synthesis efforts
Sampling	· Any sampling method
Types of evidence	· Published or unpublished academic literature, grey literature (government or institutional reports), or media reports. Slide deck presentations were included if we could identify the person giving the presentation and the date of the presentation
Outcome measures	· Reports a seroprevalence estimate (proportion of the population with detectable antibodies) · Reports the number of participants enrolled in the study (denominator) · Reports study sampling end date/week · Reports the locations at which the study took places such that they could be categorized as neighbourhood, city, state/province/territory, or country
Languages	· Any

Criteria for excluding evidence (if any met then exclude)

Characteristics	Criteria for exclusion

Population	· Non-human (e.g., *in silico*, animal, *in vitro*) · The study <u>only</u> included individuals with suspected, active, or previously diagnosed with COVID-19 using PCR, antigen testing, clinical assessment, or self-assessment · The study <u>only</u> included individuals vaccinated against SARS-CoV2
Study design	· Study designs other than cross-sectional or cohort design: case reports, case-control studies, evaluations of serological tests, study protocols
Sampling	· N/A
Types of evidence	· Multimedia sources of data (audio clips, video clips) were excluded due to the feasibility of extracting. Slide deck presentations were excluded if we could not identify the person giving the presentation and the date of the presentation · Dashboards not associated with a defined serology study
Outcome measures	· Only reports incidence or prevalence of SARS-CoV-2 antigen (as opposed to antibody) · Does not report study sampling end date/week · Does not report the number of participants included in the study (sample denominator) · Does not report the location at which the study took place
Language	· N/A

Simplified Objectives + Inclusion/Exclusion Criteria

We aimed to systematically review the evidence for the magnitude and duration of the effectiveness of (i) previous infection and (ii) hybrid immunity against multiple clinical outcomes of SARS-CoV-2 infection caused by the omicron variant. We also aimed to examine the comparative protection of hybrid immunity relative to previous infection only, vaccination only, and hybrid immunity with fewer vaccine doses.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion Criteria (all must be fulfilled):

Population

1. Humans of any age, in any geographical setting.

Exposure Group

1. Confirmed case of SARS-CoV-2 infection with or without COVID-19 vaccination.

A. SARS-CoV-2 infection will be defined as a confirmed case according to the following criteria, adapted from WHO case definitions (positive nucleic acid amplification test (NAAT) according to laboratory records or self report, positive SARS-CoV-2 antigen rapid diagnostic test (AgRDT)[a] with high accuracy according to laboratory records or self report, or a positive serology test from a lab-based assay (i.e. CLIA/ELISA) or an antibody-detecting rapid diagnostic test (Ab-RDT) with high accuracy[a]).

2. Studies will be included if they report on individuals with previously confirmed infection that have documented vaccination (partially, fully, or boosted), as defined in the randomized controlled trials for each vaccine.

A. Partial vaccination will be defined as $\geq$14 days after a single dose of Pfizer/BioNTech-Comirnaty, <7 days from the second dose for Pfizer/BioNTech-Comirnaty (BNT162b2), $\geq$14 days after a single dose of AstraZeneca-Vaxzevria, <14 days from the second dose for AstraZeneca-Vaxzevria, $\geq$14 days after a single dose of Moderna-mRNA-1273, <14 days from the second dose of Moderna-mRNA-1273, <14 days from the first dose of Janssen-Ad26.COV2.S, and $\geq$14 days after a single dose of Sinovac-CoronaVac.

B. Full vaccination will be defined as >7 days from the second dose for Pfizer/BioNTech-Comirnaty, >14 days from the first dose of Janssen-Ad26.COV2.S, >14 days from the second dose for AstraZeneca-Vaxzevria, Moderna-mRNA-1273, or Sinovac-CoronaVac.

C. Booster vaccination one will be defined as >=7 days from an additional dose after full vaccination.

D. Booster vaccination two will be defined as >=7 days from an additional dose after booster vaccination one.

Comparison Group

1. no previous vaccinations and no previously confirmed SARS-CoV-2 infection defined using WHO criteria;

2. previously confirmed SARS-CoV-2 infection defined using WHO criteria;

3. partial vaccination (defined above);

4. full vaccination (defined above);

5. booster vaccination (defined above).

Outcome

1. SARS-CoV-2 reinfection defined as a possible, probable, or confirmed reinfection case according to the following criteria, adapted from WHO case definitions.

A. Possible reinfection case will be defined as NAAT or AgRDT SARS-CoV-2 positive case with a history of a primary SARS-CoV-2 infection diagnosed by serology, with at least 60 days between the positive serology test and the subsequent positive NAAT or AgRDT.

B. Probable reinfection case will be defined as NAAT or AgRDT SARS-CoV-2 positive case with a history of a primary SARS-CoV-2 infection diagnosed by NAAT or AgRDT, with at least 90 days between the episodes. Alternatively, genomic evidence for the second episode is available and includes lineage that was not submitted to SARS-Cov-2 genomic databases at the time of first infection.

C. Confirmed reinfection case will be defined as two PCR positive episodes supported by viral genomic data from both episodes of infection revealing different Pango lineages. If viral genomic data reveal two distinct Pango lineages this will qualify as adequate evidence to confirm reinfection, regardless of the time elapsed between the two episodes.

Study Design

1. Test-negative case-control, traditional case-control, cross-sectional, cohort, non-randomized controlled trials, and randomized controlled trials.

Type of literature

1. Published peer-reviewed research articles, preprints, and grey literature in any language. We will prioritize peer-reviewed versions of articles for inclusion and analysis in instances where pre-print versions of peer-reviewed articles are available.

Exclusion Criteria (if any met then exclude):

Population

1. N/A

Exposure Group

1. No evidence of prior confirmed case. No information on the timing, brand, or dose number for the vaccination in hybrid immunity studies.

Comparison Group

1. N/A

Outcome

1. Prior infection studies not reporting the period of time between primary infection and reinfection such that determining reinfection according to the inclusion criteria is not possible.

2. Hybrid immunity studies not reporting the period of time between either the determination of primary infection or vaccination.

Study Design

1. Case reports, case series, incomplete randomized controlled trials, and review papers.

Type of literature

1. Media, news stories, and conference abstracts.

Original author protocol

[SARS-CoV-2 protective effectiveness of prior infection and hybrid immunity: a systematic review protocol](#)

2.2 Study inclusion and exclusion criteria

Table 1. Inclusion criteria	
Population	Humans of any age, in any geographical setting.

| Exposure group | Confirmed case of SARS-CoV-2 infection with or without COVID-19 vaccination.

SARS-CoV-2 infection will be defined as a **confirmed case** according to the following criteria, adapted from WHO case definitions[1] (positive nucleic acid amplification test (NAAT) according to laboratory records or self report, positive SARS-CoV-2 antigen rapid diagnostic test (AgRDT)[a] with high accuracy according to laboratory records or self report, or a positive serology test from a lab-based assay (i.e. CLIA/ELISA) or an antibody-detecting rapid diagnostic test (Ab-RDT) with high accuracy[a]).

Studies will be included if they report on individuals with previously confirmed infection that have documented vaccination (partially, fully, or boosted), as defined in the randomized controlled trials for each vaccine.

Partial vaccination will be defined as ≥14 days after a single dose of Pfizer/BioNTech-Comirnaty, <7 days from the second dose for Pfizer/BioNTech-Comirnaty (BNT162b2), ≥14 days after a single dose of AstraZeneca-Vaxzevria, <14 days from the second dose for AstraZeneca-Vaxzevria, ≥14 days after a single dose of Moderna-mRNA-1273, <14 days from the second dose of Moderna-mRNA-1273, <14 days from the first dose of Janssen-Ad26.COV2.S, and ≥14 days after a single dose of Sinovac-CoronaVac.

Full vaccination will be defined as >7 days from the second dose for Pfizer/BioNTech-Comirnaty, >14 days from the first dose of Janssen-Ad26.COV2.S, >14 days from the second dose for AstraZeneca-Vaxzevria, Moderna-mRNA-1273, or Sinovac-CoronaVac.

Booster vaccination one will be defined as >=7 days from an additional dose after full vaccination.

Booster vaccination two will be defined as >=7 days from an additional dose after booster vaccination one. |
|---|---|

Comparison group	Five comparison groups will be eligible: (1) no previous vaccinations and no previously confirmed SARS-CoV-2 infection defined using WHO criteria; (2) previously confirmed SARS-CoV-2 infection defined using WHO criteria; (3) partial vaccination (defined above); (4) full vaccination (defined above); (5) booster vaccination (defined above).
Outcome	SARS-CoV-2 reinfection defined as a possible, probable, or confirmed reinfection case according to the following criteria, adapted from WHO case definitions. **Possible reinfection case** will be defined as NAAT or AgRDT SARS-CoV-2 positive case with a history of a primary SARS-CoV-2 infection diagnosed by serology, with at least 60 days between the positive serology test and the subsequent positive NAAT or AgRDT. **Probable reinfection case** will be defined as NAAT or AgRDT SARS-CoV-2 positive case with a history of a primary SARS-CoV-2 infection diagnosed by NAAT or AgRDT, with at least 90 days between the episodes. Alternatively, genomic evidence for the second episode is available and includes lineage that was not submitted to SARS-Cov-2 genomic databases at the time of first infection. **Confirmed reinfection case** will be defined as two PCR positive episodes supported by viral genomic data from both episodes of infection revealing different Pango lineages. If viral genomic data reveal two distinct Pango lineages this will qualify as adequate evidence to confirm reinfection, regardless of the time elapsed between the two episodes.
Study design	Test-negative case-control, traditional case-control, cross-sectional, cohort, non-randomized controlled trials, and randomized controlled trials.
Type of literature	Published peer-reviewed research articles, preprints, and grey literature in any language. We will prioritize peer-reviewed versions of articles for inclusion and analysis in instances where pre-print versions of peer-reviewed articles are available.

[a]For AgRDT ≥80% sensitivity and ≥97% specificity of test, compared to NAAT, in suspected cases of infection; For AbRDT, ≥90% sensitivity (>14 days post symptom onset) and ≥97% specificity, compared to a reference lab-based test, in suspected cases of infection.

Table 2. Exclusion criteria	
Population	N/A
Exposure group	No evidence of prior confirmed case. No information on the timing, brand, or dose number for the vaccination in hybrid immunity studies.
Comparison group	N/A
Outcome	Prior infection studies not reporting the period of time between primary infection and reinfection such that determining reinfection according to the inclusion criteria is not possible. Hybrid immunity studies not reporting the period of time between either the determination of primary infection or vaccination.
Study design	Case reports, case series, incomplete randomized controlled trials, and review papers.
Type of literature	Media, news stories, and conference abstracts.

PA-Testing

Simplified Objectives + Inclusion/Exclusion Criteria

The purpose of this study was to assess the characteristics of confirmatory tests for primary aldosteronism (PA) and to interpret these in the context of study design and potential risks of bias.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion Criteria (all must be fulfilled):

1. The study is about primary aldosteronism (PA). Please note that primary hyperaldosteronism is a synonym.
2. The study examined at least one of the guideline-recommended confirmatory tests for PA. We are interested in the saline infusion test (SIT), oral salt loading test (SLT), fludrocortisone suppression test (FST), and the captopril challenge test (CCT). Please note some may have synonyms (e.g., SIT = intravenous saline suppression test [IVSS]).
3. Research articles reporting original data
4. The study is conducted in humans of any age
5. The study is in the English language
6. The performance of the confirmatory test(s) was compared with an independent reference standard A reference standard needs to be present to verify disease status. These may include: (i) clinical response to treatment (adrenalectomy and/or medical therapy), (ii) adrenal vein sampling results, (iii) histopathology, (iv) or another confirmatory test.
7. The confirmatory test was used to diagnose PA, rather than exclusively for subtyping. We are interested in knowing how confirmatory testing works for diagnosing PA, not just subtyping. If the confirmatory test was for subtyping, it is still possible to compare how many people had PA vs. non-PA.
8. The data (as published) is extractable for a 2x2 table (TP, FP, FN, TN). If the data are reported in any of the following formats, a 2x2 table can be reconstructed:(i) 2x2 table given, (ii) TP, FP, FN, TN rates given, (iii) Total number of patients (with disease) and total number of study subjects (with and without disease) is known, and corresponding sensitivity and specificity are given.

Exclusion criteria (if any met then exclude):

1. No mention about primary aldosteronism (PA)
2. Use of confirmatory tests that fall outside of guide-line recommendations
3. Non-human (e.g., in silico, animal, in vitro).
4. Conference abstracts, reviews (systematic reviews and narrative reviews), editorials, protocols, and secondary publications (data already published in another study).
5. No comparison of confirmatory test performance with a reference standard.

6. Confirmatory test was used only for subtyping, and unable to compare cases of PA vs. non-PA

7. Data not extractable for 2x2 table

Original author protocol

<u>Confirmatory Testing in Primary Aldosteronism Systematic Review Reference Sheet</u>
(Version: May 28, 2021)

<u>Primary Screen (Title/Abstract)</u>

Question 1: *Is the study about primary aldosteronism? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

  a. Yes

  b. No

  c. Unclear

Diagnosis of interest	We are interested in primary aldosteronism (PA). Please note that primary hyperaldosteronism is a synonym.

Question 2: *Does the study examine at least one of the guideline-recommended confirmatory tests for PA? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

  a. Yes

  b. No

  c. Unclear

Index test	We are interested in the saline infusion test (SIT), oral salt loading test (SLT), fludrocortisone suppression test (FST), and the captopril challenge test (CCT). Please note some may have synonyms (e.g., SIT = intravenous saline suppression test [IVSS]).

Question 3: *Is this a research study reporting original data? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

  a. Yes

  b. No

  c. Unclear

Index test	We are only interested in original studies. We will exclude conference abstracts, reviews, editorials, and protocols.

Question 4: *Is this article potentially relevant but in one of the following formats?*

    a.   Conference abstract

    b.   Systematic review

    c.   Narrative review

    d.   Secondary publication (data already published in another study)

Other reports	Although these articles will not be abstracted for our data analysis, we will still save them and later review them for background information.

Question 5: *Is this a human study? [*Exclude if "No." Include for full-text review if "Yes"]*

    a.   Yes

    b.   No

Population	We will only be considering human studies.

Question 6: *Is this an English study? [*Exclude if "No." Include for full-text review if "Yes"]*

    a.   Yes

    b.   No

Population	We will only be considering English studies.

Secondary Screen (Full-text)

Question 1: *Was the performance of the confirmatory test(s) compared with an independent reference standard? [*Exclude if "No." Include for final analysis if "Yes"]*

    a.   Yes

    b.   No

Reference test	A reference standard needs to be present to verify disease status. These may include:
	· Clinical response to treatment (adrenalectomy and/or medical therapy)
	· Adrenal vein sampling results
	· Histopathology
	· Another confirmatory test

Question 2: *Was the confirmatory test used to diagnose PA, rather than exclusively for subtyping? [*Exclude if "No." Include for final analysis if "Yes"]*

    a.   Yes: either for diagnosing PA; or if it was for subtyping, it is still possible to compare how many people had PA vs. non-PA

    b.   No: for subtyping only, and unable to compare cases of PA vs. non-PA

Diagnosis of interest	We are interested in knowing how confirmatory testing works for diagnosing PA, not just subtyping.

Question 3: *Are the data (as published) extractable for 2x2 table (TP, FP, FN, TN)? [*Exclude if "No." Include for final analysis if "Yes"]*

    a.   Yes

    b.   No

Outcome	If the data are reported in any of the following formats, a 2x2 table can be reconstructed: · 2x2 table given · TP, FP, FN, TN rates given · Total number of patients (with disease) and total number of study subjects (with and without disease) is known, and corresponding sensitivity and specificity are given

PIRD Framework:

Population = patients suspected of having primary aldosteronism (PA)

Index test = confirmatory test for PA

Reference standard = clinical response to targeted treatment (gold), adrenal vein sampling lateralization (surrogate), histopathology (surrogate), or another confirmatory test (surrogate)

Diagnosis of interest = PA

<u>PA-Outcomes</u>

Simplified Objectives and Inclusion/Exclusion Criteria

We aimed to conduct a meta-analysis to examine the clinical outcomes of surgery vs medical therapy with respect to mortality, composite major adverse cardiovascular events (MACE, and its individual components), progression to chronic kidney disease, and incident diabetes mellitus.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

<u>Inclusion Criteria (all must be fulfilled):</u>

1. The study is about primary aldosteronism (PA). Please note that primary hyperaldosteronism is a synonym.
2. Reports surgery (adrenalectomy) or medication (mineralocorticoid receptor antagonists, spironolactone, eplerenone) treatment for primary aldosteronism.
3. Reports mortality, MACE, ACS, stroke, arrhythmia, heart failure, chronic kidney disease, and/or incident diabetes clinical outcomes after treatment.
4. Research articles reporting original data.
5. Randomized clinical trials, cohort studies, and cross-sectional studies.
6. The study is conducted in humans of any age.
7. The study is in the English language.

<u>Exclusion criteria (if any met then exclude):</u>

1. Does not report about primary aldosteronism (PA).
2. Does not report treatment for primary aldosteronism.
3. Does not report mortality, major adverse cardiovascular events (MACE), acute coronary syndrome (ACS), stroke, arrhythmia, heart failure, chronic kidney disease, and/or incident diabetes clinical outcomes after treatment.
4. Non-human (e.g., in silico, animal, in vitro).
5. Conference abstracts, case reports, case series, reviews (systematic reviews and narrative reviews), editorials, protocols, and secondary publications (data already published in another study).

Original author protocol

Primary Aldosteronism Treatment Response: Systematic Review Reference Sheet
(Version: June 1, 2022 revised)

Primary Screen (Title/Abstract)

Question 1: *Is the study about primary aldosteronism? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

      a.   Yes

      b.   No

      c.   Unclear

Diagnosis of interest	We are interested in primary aldosteronism (PA). Please note that primary hyperaldosteronism is a synonym.

Question 2: *Does this study describe treatment for primary aldosteronism? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

      a.   Yes

      b.   No

      c.   Unclear

Intervention (exposure) of interest	We are interested in treatment outcomes for primary aldosteronism (PA), comparing surgery (adrenalectomy) vs. medications (mineralocorticoid receptor antagonists, spironolactone, eplerenone).

Question 3: *Does this study describe hard clinical outcomes? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

      a.   Yes

      b.   No

      c.   Unclear

Outcomes of interest	We are interested in hard clinical outcomes for primary aldosteronism (PA), such as mortality, MACE, ACS, stroke, arrhythmia, heart failure, chronic kidney disease, and/or incident diabetes.

Question 4: *Is this a research study reporting original data? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

     a. Yes

     b. No

     c. Unclear

Design	We are only interested in original studies. We will exclude conference abstracts, reviews, editorials, and protocols.

Question 5: *Is this article potentially relevant but in one of the following formats?*

     a. Conference abstract

     b. Systematic review

     c. Narrative review

     d. Secondary publication (data already published in another study)

Other reports	Although these articles will not be abstracted for our data analysis, we will still save them and later review them for background information.

Question 6: *Is this a human study? [*Exclude if "No." Include for full-text review if "Yes"]*

     a. Yes

     b. No

Population	We will only be considering human studies.

Question 7: *Is this an English study? [*Exclude if "No." Include for full-text review if "Yes"]*

     a. Yes

     b. No

Language	We will only be considering English studies.

Secondary Screen (Full-text)

Question 1: *Does the study report on an outcome of interest? [*Exclude if "No." Include for final analysis if "Yes"]*

a. Yes

b. No

Outcome	The study must report on an outcome of interest with a quantitative metric (either as a summary statistic, like a HR, RR, or OR; or as a survival/cumulative incidence curve). These may include: · Mortality, all-cause death · Cardiovascular death · Major adverse cardiovascular events / composite cardiovascular events · Acute coronary syndrome / myocardial infarction / unstable angina / coronary heart disease · Coronary revascularization (PCI, CABG) · Stroke (stroke, TIA) · Arrhythmias (atrial fibrillation, atrial flutter, ventricular fibrillation, ventricular tachycardia) · Heart failure / heart failure hospitalization · Diabetes · Chronic kidney disease

Question 2: *Does the study directly compare medication (e.g., mineralocorticoid receptor antagonist) vs. surgery for patients with PA, and additionally stratify outcomes according to the treatment received? [*Exclude if "No." Include for final analysis if "Yes"]*

    a. Yes: it is possible to compare outcomes for PA patients who had surgery vs. medications

    b. No: unable to compare outcomes for surgery vs. medications (e.g., because all patients with PA are grouped together)

Exposure of interest	We are interested in knowing how treatment outcomes differ between surgery and medications (not just the natural history of PA).

Question 3: *Does this study report a hard clinical outcome that can be extracted according to treatment received? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

    a. Yes

    b. No

    c. Unclear

Outcomes of interest	We are interested in hard clinical outcomes for primary aldosteronism (PA), such as mortality, MACE, ACS, stroke, arrhythmia, heart failure, chronic kidney disease, and/or incident diabetes.

Question 4: *Is this a research study reporting original data? [*Exclude if "No." Include for full-text review if "Yes" or "Unclear"]*

        a.   Yes

        b.   No

        c.   Unclear

Design	We are only interested in original studies. We will exclude conference abstracts, reviews, editorials, and protocols.

<u>**PECOD Framework**</u>:

The *population* of interest are patients with PA.

The *exposure/intervention* of interest is surgical adrenalectomy and the *comparator* is medical treatment with a mineralocorticoid receptor antagonist (e.g., spironolactone, eplerenone, etc).

The *primary outcome* is all-cause mortality. *Secondary outcomes* include other commonly reported clinical events include incident major adverse cardiovascular events (and its individual components, such as myocardial infarction, stroke, revascularization, arrhythmia), congestive heart failure, atrial fibrillation, chronic kidney disease, and diabetes mellitus.

The study *designs* that will be considered include randomized controlled trials and observational studies.

<u>SVCF</u>

Simplified Objectives + Inclusion/Exclusion Criteria

Objectives: To evaluate the association of low SVC flow, diagnosed in the first 48 hours after birth echocardiography, with neurological morbidity and mortality, among very preterm neonates.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

<u>Inclusion Criteria (all must be fulfilled):</u>

1. Preterm infants <32 weeks gestational age who had echocardiography done within the first 48 hours after birth with evaluation of SVC flow
A. Prognostic factor: Low SVC flow identified by Doppler assessment during echocardiography performed in the first 48 hours after birth.
B. Comparison: Normal SVC Flow
2. Reported outcomes including intraventricular hemorrhage (IVH), presence of periventricular leukomalacia (PVL), all-cause mortality before discharge from neonatal intensive care unit (NICU), neurodevelopmental impairment in early childhood or any diagnosed Cerebral Palsy, visual and/or hearing deficits, or necrotizing enterocolitis (NEC)
A. Any grade IVH diagnosed in the first 7 days after birth by cranial ultrasonography
B. Severe IVH (defined as stage 3 or higher according to Papile's classification 14) diagnosed in the first 7 days after birth
3. Randomized controlled trials, cohort or case-control studies

<u>Exclusion criteria (if any met then exclude):</u>

1. Non-human (e.g., in silico, animal, in vitro)
2. Cross-sectional studies, narrative reviews, case series or case reports

Original author protocol

<u>Association of early-life low superior vena cava flow among preterm neonates and death or cerebral haemorrhage: a systematic review and meta-analysis: Protocol</u>

The objective of this study is to systematically review and meta—analyse the association of low SVC flow during the transitional period among preterm neonates < 32 weeks GA with mortality and adverse neurological morbidity..

PICO/PFO outline-

<u>*Population*</u> – Preterm infants <32 weeks gestational age

<u>*Prognostic factor-*</u> Low SVC flow identified by Doppler assessment during echocardiography performed in the first 48 hours after birth (measured as ml/kg/min)

<u>*Comparison-*</u> Patients with normal SVC flow

<u>*Outcomes-*</u>
Primary- **Any grade IVH** diagnosed in the first 7 days of life by cranial ultrasonography
Secondary-

- Severe IVH (defined as stage 3 or higher according to Papile's classification [14]) diagnosed in the first 7 days of life,
- Presence of PVL (diagnosed by 28 days of life)
- Mortality within the neonatal period (defined as the first 28 days of life)
- Neurodevelopmental impairment in early childhood (Defined as a composite outcome of any of the following: Cerebral palsy with Gross Motor Function Classification System score >=1 or Bayley-III motor composite <85; Bayley cognitive composite < 85; Bayley language composite <85; Any sensorineural/mixed hearing loss; Any unilateral or bilateral visual impairment)[15]
- Necrotizing Enterocolitis (NEC) compared between low and normal SVC flow groups

METHODOLOGY

This systematic review and meta-analysis will be conducted according to the PRISMA guidelines and Cochrane methodology.

Eligibility Criteria

Randomized controlled trials, cohort or case-control studies that evaluated the following population characteristics will be included for this study:
Preterm infants <32 weeks gestational age who had echocardiography done within the first 48 hours after birth with evaluation of SVC flow. Studies must have compared any of the above stated clinical

64

outcomes among preterm neonates with low vs. normal SVC flow to be considered for inclusion. Studies will be included only if they exclusively include human subjects and there will be no language limitations.

Exclusion criteria: Cross-sectional studies, narrative reviews, case series or case reports on this topic will be excluded. We will also exclude any animal studies. Any identified studies without any available full text from the included databases or the original authors will also be excluded.

Sepsis

Simplified Objectives and Inclusion/Exclusion Criteria

Objectives: To assess comparative effectiveness and safety of fludrocortisone plus hydrocortisone, hydrocortisone alone, and placebo/usual care in adults with septic shock.

Inclusion Criteria (all must be fulfilled):

1. Randomized control trials (RCTs)

2. Patients > 16 years with septic shock

A. Septic shock was defined as sepsis and use of at least one vasopressor

3. Eligible groups included any of the following:

A. Hydrocortisone

B. Hydrocortisone-fludrocortisone

C. Placebo

D. Usual care

Exclusion criteria (if any met then exclude):

1. Studies that were not randomized control trials (RCTs), including reviews, observational studies, case reports, case series

2. Trials evaluating hydrocortisone as part of ascorbic acid- or thiamine-based 'metabolic resuscitation'

Original author protocol

METHODS

Protocol and registration

The protocol for this review was registered on PROSPERO (CRD42023395927) on 02/15/2023 and modified on 03/21/2023 to include patients ≥16 years (previously ≥18 years). We followed the Preferred Reporting Items for Systematic Reviews incorporating Network Meta-Analyses guidelines (17) to report this study.

Eligibility criteria

We included RCTs that assigned individuals aged ≥16 years with septic shock (sepsis and use of at least one vasopressor) to receive hydrocortisone alone, hydrocortisone-fludrocortisone, placebo or usual care. Placebo and usual care were considered interchangeable in the main analysis, but lack of placebo control was considered in the risk of bias assessment. Trials evaluating hydrocortisone as part of ascorbic acid- or thiamine-based 'metabolic resuscitation' (18) were excluded. We included published peer-reviewed trials regardless of language if abstracts were reported in English. Reports in non-English languages were translated. Studies were limited to those published starting 8/1991 when the first consensus definition of sepsis was established by the Society for Critical Care Medicine and American College of Chest Physicians (19).

Spinal

Simplified Objectives and Inclusion/Exclusion Criteria

In an effort to prevent intraoperative neurological injury during spine surgery, the use of intraoperative neurophysiological monitoring (IONM) has increased significantly in recent years. Using IONM, spinal cord function can be evaluated intraoperatively by recording signals from specific nerve roots, motor tracts, and sensory tracts. We performed a systematic review and meta-analysis of diagnostic test accuracy (DTA) studies to evaluate the efficacy of IONM among patients undergoing spine surgery for any indication.

Inclusion Criteria (all must be fulfilled):

Population

1. Adolescents (≥12 years to <18 years old) or adults (≥18 years) undergoing any type of spine surgery for any indication or spine-related pathology (including trauma-related pathology, conus injuries, cauda equina injuries)

A. Subpopulations/groups of interest: Spinal deformity (e.g., scoliosis), Intra-dural tumors vs. extra-dural lesions, Degenerative spine lesions

Intervention

1. Neuromonitoring including somatosensory evoked potentials (SSEP), motor evoked potential (MEP), electromyography (EMG) and multimodal monitoring

Outcomes

1. True positives (a change in monitoring corresponding to a change in postoperative neurologic status)

2. False positive (a change in monitoring corresponding to no change in postoperative neurologic status)

3. False negative (no change in monitoring, but a change in postoperative neurologic status

4. True negative (no change in monitoring corresponding to no change in postoperative neurologic status)

Reference standard

1. Postoperative Neurological change/recovery (based on validated measures):

A. AIS grade

B. Motor score

C. Frankel grade

Timing

1. Immediate post-operative period

Study design

1. Randomized control trials (RCTs) or comparative observational studies (comparative cohorts, case control studies)

2. Case series will be included if greater than 10 patients

Exclusion criteria (if any met then exclude):

Population

1. Patients < 11 years old

2. Patients with new post-operative compression (e.g., hematoma, abscess)

3. Patients with pedicle screw breach

4. Patients with neurologic deficits due to cranial pathology (e.g., stroke)

Reference standard

1. Pedicle screw breach

Study design

1. Case series with <10 patients, animal studies, abstracts, editorials, letters, single reports from multicenter trials, white papers, narrative reviews, proceedings/abstracts from meetings

Original author protocol

<u>Criteria for Inclusion/Exclusion of Studies in the Review</u>
The criteria for inclusion and exclusion of studies for this systematic review will be based on the criteria specified a priori for population, interventions, outcomes, reference standard, timing, and settings/studies (PICOTS) listed in Table 1.

Table 1. Inclusion and exclusion criteria: population, interventions, comparators, outcomes, timing, and study designs

	Inclusion	Exclusion
Population	• Adolescents (≥12 years to <18 years old) or adults (≥18 years) undergoing any type of spine surgery for any indication or spine-related pathology (including trauma-related pathology, conus injuries, cauda equina injuries Subpopulations/groups of interest • Spinal deformity (e.g., scoliosis) • Intra-dural tumors vs. extra-dural lesions • Degenerative spine lesions	• Patients <11 years old • Patients with new post-operative compression (e.g., hematoma, abscess) • Patients with pedicle screw breach • Patients with neurologic deficits due to cranial pathology (e.g., stroke)
Intervention	• Neuromonitoring including somatosensory evoked potentials (SSEP), motor evoked potential (MEP), electromyography (EMG) and multimodal monitoring	
Outcomes	• True positives (a change in monitoring corresponding to a change in postoperative neurologic status) • False positive (a change in monitoring corresponding to no change in postoperative neurologic status) • False negative (no change in monitoring, but a change in postoperative neurologic status • True negative (no change in	

	monitoring corresponding to no change in postoperative neurologic status)	

Reference Standard	● Postoperative Neurological change/recovery (based on validated measures)
	  ○ AIS grade
	  ○ Motor score
	  ○ Frankel grade
		● Pedicle screw breach

Timing	● Immediate post-operative period	
Study Design	The review focus on the evidence at least risk for bias ● RCTs or Comparative observational studies (comparative cohorts, case control studies) ● Case series will be included if comparative studies are not available	● Case series with ≤10 patients ● Animal studies ● Abstracts, editorials, letters ● Duplicate publications of the same study that do not report on different outcomes ● Single reports from multicenter trials ● White papers ● Narrative reviews ● Proceedings/abstracts from meetings ● Articles identified as preliminary reports when results are published in later versions

<u>Calcium</u>

Simplified Objectives and Inclusion/Exclusion Criteria

Objectives: To compare the outcomes of routine calcium administration to no calcium administration for cardiac arrest in adults or children

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

<u>Inclusion Criteria (all must be fulfilled):</u>
1. Adults and children in any setting (in-hospital or out-of-hospital) with cardiac arrest
2. Reports administration of calcium (intravenous or intraosseous) during cardiac arrest as compared to no administration of calcium during cardiac arrest. Other eligible terms for calcium include calcium carbonate, calcium citrate, and calcium gluconate.
3. Reports relevant outcomes, including return of spontaneous circulation (ROSC), health-related quality of life, survival and survival with favourable neurologic outcome to hospital discharge, 30 days or longer.
4. Randomised controlled trials (RCTs) and non-randomized studies (non-randomized controlled trials, interrupted time series, controlled before-and-after studies, cohort studies) with a control group were eligible for inclusion.

<u>Exclusion criteria (if any met then exclude):</u>
1. No mention of cardiac arrest
2. No mention of calcium administration during cardiac arrest, or comparison between calcium administration and no calcium administration during cardiac arrest.
4. No mention of relevant outcomes
5. Ecological studies, case series, case reports, reviews, abstracts, editorials, comments, letters to the editor, and unpublished studies

Original author protocol

The study question was framed using the PICOST (Population, Intervention, Comparison, Outcome, Study Design, Time frame) format: in adults and children in any setting (in-hospital or out-of-hospital) with cardiac arrest (P), does administration of calcium (intravenous or intraosseous) during cardiac arrest (I) as compared to no administration of calcium during cardiac arrest (C) improve clinical outcomes (O).

Relevant outcomes were prioritised by the ILCOR Advanced Life Support and Paediatric Life Support Task Forces and based on the available literature. We included return of spontaneous circulation (ROSC), health-related quality of life, survival and survival with favourable neurologic outcome to hospital discharge, 30 days or longer.

Randomised controlled trials (RCTs) and non-randomized studies (non-randomized controlled trials, interrupted time series, controlled before-and-after studies, cohort studies) with a control group were eligible for inclusion. Ecological studies, case series, case reports, reviews, abstracts, editorials, comments, letters to the editor, and unpublished studies were excluded. All years and all languages were included if there is an English abstract.

Simplified Objectives and Inclusion/Exclusion Criteria

A systematic review and meta-analysis of observational studies was conducted to examine the association between immediate response (improved oxygenation ≤6 h) compared to non-response, and all-cause mortality among preterm infants <34 weeks gestational age without congenital anomalies or genetic disorders who received inhaled nitric oxide (iNO) treatment.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion Criteria (all must be fulfilled):
1. Preterm infants born < 34 weeks GA free from congenital anomalies or genetic disorders treated with inhaled Nitric Oxide (iNO) at any point during NICU or hospital admission
2. Comparative studies that divided the cohort based on iNO responsiveness, defined by an absolute reduction in FiO2 or a synonymous measure of oxygenation within ≤6 h of treatment initiation, were included.
3. Reported pre-discharge mortality outcomes, which was defined as death from any cause occurring prior to hospital discharge. Additional eligible outcomes included severe neurological injury, defined as intraventricular hemorrhage (IVH) ≥ grade 3 or periventricular leukomalacia (PVL), overall and among survivors, bronchopulmonary dysplasia (BPD) diagnosed at 36 weeks postmenstrual age (PMA) or at hospital discharge if <36 weeks PMA in survivors only, and neurodevelopmental impairment at a minimum of 18–24 months post-discharge.
4. Prospective and retrospective cohort studies, any language or publication date

Exclusion criteria (if any met then exclude):
1. Preterm infants with congenital anomalies or genetic disorders
2. No mention of preterm infants treated with inhaled Nitric Oxide (iNO)
3. No evaluation of iNO responsiveness
4. No mention of pre-discharge mortality, severe neurological injury, bronchopulmonary dysplasia (BPD), or neurodevelopmental impairment clinical outcomes

5. Case series, grey literature, conference abstracts, editorials, letters to the editor and commentary, reviews (systematic reviews and narrative reviews)

Original author protocol

2.1. Eligibility criteria

Prospective and retrospective cohort studies including preterm infants born <34 weeks of gestational age (GA) without congenital anomalies or genetic disorders, treated with iNO during hospital admission were included. There were no restrictions on language or publication dates. Grey literature and conference abstracts were excluded.

2.2. Exposure

Comparative studies that divided the cohort based on iNO responsiveness, defined by an absolute reduction in FiO2 or a synonymous measure of oxygenation within ≤6 h of treatment initiation, were included.

2.3. Outcomes

The primary outcome was pre-discharge mortality, which was defined as death from any cause occurring prior to hospital discharge. Additional outcomes included severe neurological injury, defined as intraventricular hemorrhage (IVH) ≥ grade 3 or periventricular leukomalacia (PVL), overall and among survivors, bronchopulmonary dysplasia (BPD) diagnosed at 36 weeks postmenstrual age (PMA) or at hospital discharge if <36 weeks PMA in survivors only, and neurodevelopmental impairment at a minimum of 18–24 months post-discharge.

Study design
Included: Prospective and retrospective cohort studies, any language or publication date
Excluded: Case series, grey literature, conference abstracts, editorials, letters to the editor and commentary

Protocol was registered PROSPERO, CRD42023408759

<u>Meds-HA</u>

Simplified Objectives and Inclusion/Exclusion Criteria

The objectives of this study were to systematically review the evidence to identify medications that affect hospital admissions and prioritize therapies for quality improvement by assessing the quality of evidence and cross-referencing the findings with clinical guidelines.

<u>Inclusion Criteria (all must be fulfilled):</u>

1. We included systematic reviews of RCTs published in English that examined the effect of any medication on emergency hospital admissions in adults (16 years or older).

2. We included reviews that searched two or more electronic databases and assessed and reported the quality of included studies. We defined a medication as any administered chemical or biological product.

3. We included only reviews that reported at least one meta-analysed effect estimate for emergency hospital admissions that was not part of a composite measure.

4. We defined an emergency hospital admission as an unanticipated admission or readmission to hospital that occurred at short notice because of a perceived need for immediate health care. We did not consider admission to an emergency department or an observational unit to be a hospital admission.

<u>Exclusion criteria (if any met then exclude):</u>

1. We excluded studies reporting only scheduled or elective hospital admissions.

2. We excluded studies if they were not systematic reviews with meta-analysed producing pooled estimates

3. We excluded studies if there were not systematic reviews that included randomized controlled trials

4. We excluded superseded Cochrane reviews (Cochrane reviews that have been replaced by newer Cochrane reviews on the same topic).

5. We excluded non-Cochrane reviews if all the randomized controlled trial data on hospital admissions were included in a more recent systematic review of the same intervention and patients.

Original author protocol

INCLUDE

We included systematic reviews of RCTs published in English that examined the effect of any medication on emergency hospital admissions in adults (16 years or older).

We included reviews that searched two or more electronic databases and assessed and reported the quality of included studies. We defined a medication as any administered chemical or biological product.

We included only reviews that reported at least one meta-analysed effect estimate for emergency hospital admissions that was not part of a composite measure

We defined an emergency hospital admission as an unanticipated admission or readmission to hospital that occurred at short notice because of a perceived need for immediate health care. We did not consider admission to an emergency department or an observational unit to be a hospital admission.

EXCLUDE

We excluded studies reporting only scheduled or elective hospital admissions

We excluded studies if they were not systematic reviews with meta-analysed producing pooled estimates

We excluded studies if there were not systematic reviews that included randomized controlled trials

We excluded superseded Cochrane reviews (Cochrane reviews that have been replaced by newer Cochrane reviews on the same topic)

We excluded non-Cochrane reviews if all the randomized controlled trial data on hospital admissions were included in a more recent systematic review of the same intervention and patients.

When two reviews reported identical underlying clinical trial data, we selected the review that reported more detailed information, as judged by two of the authors through discussion and consensus.

Supplementary Note 2: Example Abstract ScreenPrompt

A full example of our Abstract ScreenPrompt prompt for the SeroTracker SR is provided below.

Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion Criteria (all must be fulfilled):
Population
1. Humans of any age
2. Including COVID-19 antigen positive persons and those with suspected disease if not deliberately sampled.
Study design
1. Sero-surveys – defined as the collection and testing of serum (or proxy such as oral fluid) specimens from a sample of a defined population over a specified period of time to estimate the prevalence of antibodies against SARS-CoV-2 as an indicator of immunity
2. Cross-sectional, repeated cross sectional, and cohort study designs, with serology measurements at single time points or repeated at multiple time points
Special design
1. Include systematic reviews and meta-analysis of seroprevalence studies for the purpose of tracking evidence synthesis efforts
Sampling
1. Any sampling method
Types of evidence
1. Published or unpublished academic literature, grey literature (government or institutional reports), or media reports. Slide deck presentations were included if we could identify the person giving the presentation and the date of the presentation
Outcome measures
1. Reports a seroprevalence estimate (proportion of the population with detectable antibodies)
2. Reports the number of participants enrolled in the study (denominator)
3. Reports study sampling end date/week
4. Reports the locations at which the study took places such that they could be categorized as neighbourhood, city, state/province/territory, or country

Exclusion Criteria (if any met then exclude):
Population
1. Non-human (e.g., in silico, animal, in vitro)

2. The study only included individuals with suspected, active, or previously diagnosed with COVID-19 using PCR, antigen testing, clinical assessment, or self-assessment
3. The study only included individuals vaccinated against SARS-CoV2
Study design
1. Study designs other than cross-sectional or cohort design: case reports, case-control studies, evaluations of serological tests, study protocols
Types of evidence
1. Multimedia sources of data (audio clips, video clips) were excluded due to the feasibility of extracting. Slide deck presentations were excluded if we could not identify the person giving the presentation and the date of the presentation
2. Dashboards not associated with a defined serology study
Outcome measures
1. Only reports incidence or prevalence of SARS-CoV-2 antigen (as opposed to antibody)
2. Does not report study sampling end date/week
3. Does not report the number of participants included in the study (sample denominator)
4. Does not report the location at which the study took place

{ABSTRACT CONTENT}

Instructions
We now assess whether the paper should be included in the systematic review by evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.

Studies that may not fully align with the primary focus of our inclusion criteria but provide data or insights potentially relevant to our review deserve thoughtful consideration. Given the nature of abstracts as concise summaries of comprehensive research, some degree of interpretation is necessary.

Our aim should be to inclusively screen abstracts, ensuring broad coverage of pertinent studies while filtering out those that are clearly irrelevant. We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.

Supplementary Note 3: Example ISO-ScreenPrompt

A full example of our ISO-ScreenPrompt prompt for the SeroTracker SR is provided below:

Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion Criteria (all must be fulfilled):
1. Humans of any age
2. Including COVID-19 antigen positive persons and those with suspected disease if not deliberately sampled.
3. Sero-surveys – defined as the collection and testing of serum (or proxy such as oral fluid) specimens from a sample of a defined population over a specified period of time to estimate the prevalence of antibodies against SARS-CoV-2 as an indicator of immunity
4. Cross-sectional, repeated cross sectional, and cohort study designs, with serology measurements at single time points or repeated at multiple time points
5. Include systematic reviews and meta-analysis of seroprevalence studies for the purpose of tracking evidence synthesis efforts
Sampling
6. Any sampling method
7. Published or unpublished academic literature, grey literature (government or institutional reports), or media reports. Slide deck presentations were included if we could identify the person giving the presentation and the date of the presentation
8. Reports a seroprevalence estimate (proportion of the population with detectable antibodies)
9. Reports the number of participants enrolled in the study (denominator)
10. Reports study sampling end date/week
11. Reports the locations at which the study took places such that they could be categorized as neighbourhood, city, state/province/territory, or country

Exclusion Criteria (if any met then exclude):
1. Non-human (e.g., in silico, animal, in vitro)
2. The study only included individuals with suspected, active, or previously diagnosed with COVID-19 using PCR, antigen testing, clinical assessment, or self-assessment
3. The study only included individuals vaccinated against SARS-CoV2
4. Study designs other than cross-sectional or cohort design: case reports, case-control studies, evaluations of serological tests, study protocols

5. Multimedia sources of data (audio clips, video clips) were excluded due to the feasibility of extracting. Slide deck presentations were excluded if we could not identify the person giving the presentation and the date of the presentation
6. Dashboards not associated with a defined serology study
7. Only reports incidence or prevalence of SARS-CoV-2 antigen (as opposed to antibody)
8. Does not report study sampling end date/week
9. Does not report the number of participants included in the study (sample denominator)
10. Does not report the location at which the study took place

Instructions
We now assess whether the paper should be included in the systematic review by rigorously evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.

We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.

{FULL-TEXT CONTENT}

Our systematic review is governed by the following objectives: (i) describe the global prevalence of SARS-CoV-2 antibodies based on serosurveys; (ii) detect variations in seroprevalence arising from study design and geographic factors; (iii) identify populations at high risk for SARS-CoV-2 infection; and (iv) evaluate the extent to which surveillance based on detection of acute infection underestimates the spread of the pandemic.

The following is an excerpt of two sets of criteria. A study is considered included if it meets all the inclusion criteria. If a study meets any of the exclusion criteria, it should be excluded. Here are the two sets of criteria:

Inclusion Criteria (all must be fulfilled):
1. Humans of any age
2. Including COVID-19 antigen positive persons and those with suspected disease if not deliberately sampled.
3. Sero-surveys – defined as the collection and testing of serum (or proxy such as oral fluid) specimens from a sample of a defined population over a specified period of time to estimate the prevalence of antibodies against SARS-CoV-2 as an indicator of immunity
4. Cross-sectional, repeated cross sectional, and cohort study designs, with serology measurements at single time points or repeated at multiple time points
5. Include systematic reviews and meta-analysis of seroprevalence studies for the purpose of tracking evidence synthesis efforts
Sampling
6. Any sampling method

7. Published or unpublished academic literature, grey literature (government or institutional reports), or media reports. Slide deck presentations were included if we could identify the person giving the presentation and the date of the presentation

8. Reports a seroprevalence estimate (proportion of the population with detectable antibodies)

9. Reports the number of participants enrolled in the study (denominator)

10. Reports study sampling end date/week

11. Reports the locations at which the study took places such that they could be categorized as neighbourhood, city, state/province/territory, or country

Exclusion Criteria (if any met then exclude):

1. Non-human (e.g., in silico, animal, in vitro)

2. The study only included individuals with suspected, active, or previously diagnosed with COVID-19 using PCR, antigen testing, clinical assessment, or self-assessment

3. The study only included individuals vaccinated against SARS-CoV2

4. Study designs other than cross-sectional or cohort design: case reports, case-control studies, evaluations of serological tests, study protocols

5. Multimedia sources of data (audio clips, video clips) were excluded due to the feasibility of extracting. Slide deck presentations were excluded if we could not identify the person giving the presentation and the date of the presentation

6. Dashboards not associated with a defined serology study

7. Only reports incidence or prevalence of SARS-CoV-2 antigen (as opposed to antibody)

8. Does not report study sampling end date/week

9. Does not report the number of participants included in the study (sample denominator)

10. Does not report the location at which the study took place

Instructions

We now assess whether the paper should be included in the systematic review by rigorously evaluating it against each and every predefined inclusion and exclusion criterion. First, we will reflect on how we will decide whether a paper should be included or excluded. Then, we will think step by step for each criteria, giving reasons for why they are met or not met.

We will conclude by outputting (on the very last line) 'XXX' if the paper warrants exclusion, or 'YYY' if inclusion is advised or uncertainty persists. We must output either 'XXX' or 'YYY'.

General LLM Terminology:

Classifier Model: machine learning models designed to categorize input data into predefined classes or labels. These models work by identifying patterns and features in the input data and mapping them to specific categories. In the context of systematic reviews, a classifier model could be used to determine whether an article meets inclusion criteria ("include") or should be excluded ("exclude"), based on its abstract or full text. Classifier models are foundational in tasks involving categorization, prioritization, and decision-making.

Large Language Model (LLM): An advanced type of artificial intelligence typically based on deep learning architectures like transformers, trained on a massive corpora of textual data. These models, such as GPT or similar frameworks, are designed to perform a wide range of natural language processing (NLP) tasks, including text generation, classification, and summarization.

Prompting: refers to the practice of providing structured input instructions or queries to large language models (LLMs) to elicit specific responses.
- In zero-shot prompting, the LLM is given only the task description or a direct question without any examples to guide its response. The model must rely entirely on its pre-trained knowledge and reasoning capabilities to generate an answer.
- Few-shot prompting involves providing the LLM with a small number of examples (typically 1-5) within the prompt to illustrate the task. These examples serve as context for the model, helping it generate more precise or task-specific responses.
- Chain-of-thought prompting instructs LLMs to think step-by-step. In a zero-shot setting, LLMs generate 'unstructured freeform reasoning', where the step by step reasoning does not rely on rigid data structures.

Generalist foundation models: large-scale, pre-trained AI systems developed on diverse, multi-domain datasets. These models, such as GPT or BERT, form the basis for a wide range of downstream tasks without being domain-specific at their core.

Fine-tuning: the process of adapting a pre-trained machine learning model, such as a large language model (LLM) or neural network, to perform a specific task or operate in a specialized domain. This is achieved by training the model further on a smaller, task-specific dataset while leveraging the foundational knowledge it acquired during pre-training. Fine-tuning allows the model to retain general capabilities while improving performance on the target task.

BERT fine-tuning: involves adapting the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model for specific downstream tasks through additional domain-specific training. Unlike training from scratch, fine-tuning leverages BERT's robust contextual language understanding, enabling precise performance on specialized tasks, such as identifying relevant publications in a systematic review or classifying biomedical text.

Model Infrastructure:

Tokens: Tokens are the basic building blocks of the input and output processed by the model. They can represent whole words, subwords, or even individual characters, depending on the complexity of the text.

- For example, the sentence "Systematic reviews are essential" might be broken into the tokens: ["Systematic", " reviews", " are", " essential"].

Input tokens: tokens derived from the text you provide to the model (e.g., a query or abstract for screening). These are counted toward the token limit.

Output tokens: tokens generated by the model as a response. The combined total of input and output tokens must stay within the maximum tokens limit.

Maximum tokens: This parameter defines the total limit for output tokens in a single API call. For example, if the maximum tokens are set to 4096, the model can generate up to 4096 output tokens.

Evaluation token: To evaluate the model responses, we required an output containing an 'evaluation token': either "XXX" (exclude) or "YYY" (include).

Unparseable requests: For each request sent to the API, if a response was interrupted by a technical error (rate limit, disconnect, timeout) or was missing an 'evaluation token', we retried the instance up to three times. If all three attempts failed, we deemed the request as 'unparseable' and disqualified the request from the total count, occasionally reducing the set by 1-2 articles. However, after shifting to the openAI Batch API for our re-analysis greatly reduced technical errors and rarely resulted in missing "evaluation tokens".

Seed parameter: The seed parameter sets the initial state of the model's random number generator, ensuring that the same input produces identical outputs across multiple runs. This is useful for reproducibility in academic settings, where consistent results are critical for validation and comparisons.

API call: An API call is a request made by one software application to another through an Application Programming Interface (API) to retrieve or send information or perform a specific task. In our context, our prompt (containing the article of interest) is sent to an LLM API to generate an output (decision to include or exclude).

Model settings

Temperature: Controls the randomness of the model's output. A higher temperature allows for more diverse and creative responses, as it increases the probability of less common words being selected. Lower temperatures produce more focused and deterministic outputs by prioritizing highly probable responses.

85

Top_p: also known as nucleus sampling, adjusts the diversity of the output by limiting the selection of tokens (words or symbols) to a subset of the most likely options.

Frequency_penalty: Frequency penalty discourages the model from repeating the same tokens too often. By applying a penalty based on how frequently a token has already been used in the text, it encourages more varied responses.

Presence_penalty: Presence penalty modifies the model's tendency to include specific tokens that have already appeared in the text. Unlike frequency penalty, which focuses on repetition count, presence penalty discourages reuse altogether, ensuring diverse content generation.

Misc
Balanced train: The training dataset used to train or fine-tune a machine learning model has an equal or proportional representation of all relevant classes or categories. For example, the 'ST train sample' had 200 'included' and 200 'excluded' citations.

Supplementary References

1. Bobrovitz N, Arora RK, Cao C, Boucher E, Liu M, Donnici C, et al. Global seroprevalence of SARS-CoV-2 antibodies: A systematic review and meta-analysis. Khudyakov YE, editor. PLOS ONE. 2021 Jun 23;16(6):e0252617.

2. Bobrovitz N, Ware H, Ma X, Li Z, Hosseini R, Cao C, et al. Protective effectiveness of previous SARS-CoV-2 infection and hybrid immunity against the omicron variant and severe disease: a systematic review and meta-regression. Lancet Infect Dis. 2023 May;23(5):556–67.

3. Samnani S, Cenzer I, Kline GA, Lee SJ, Hundemer GL, McClurg C, et al. Time to Benefit of Surgery vs Targeted Medical Therapy for Patients With Primary Aldosteronism: A Meta-analysis. J Clin Endocrinol Metab. 2024 Feb 20;109(3):e1280–9.

4. Leung AA, Symonds CJ, Hundemer GL, Ronksley PE, Lorenzetti DL, Pasieka JL, et al. Performance of Confirmatory Tests for Diagnosing Primary Aldosteronism: a Systematic Review and Meta-Analysis. Hypertension. 2022 Aug;79(8):1835–44.

5. Teja B, Berube M, Pereira TV, Law AC, Schanock C, Pang B, et al. Effectiveness of Fludrocortisone Plus Hydrocortisone versus Hydrocortisone Alone in Septic Shock: A Systematic Review and Network Meta-Analysis of Randomized Controlled Trials. Am J Respir Crit Care Med. 2024 May 15;209(10):1219–28.

6. Alvi MA, Kwon BK, Hejrati N, Tetreault LA, Evaniew N, Skelly AC, et al. Accuracy of Intraoperative Neuromonitoring in the Diagnosis of Intraoperative Neurological Decline in the Setting of Spinal Surgery—A Systematic Review and Meta-Analysis. Glob Spine J. 2024 Mar;14(3_suppl):105S-149S.

7. Hsu CH, Couper K, Nix T, Drennan I, Reynolds J, Kleinman M, et al. Calcium during cardiac arrest: A systematic review. Resusc Plus. 2023 Jun;14:100379.

8. Baczynski M, Jasani B, De Castro C, Dani C, Subhedar NV, Chandrasekharan P, et al. Association between immediate oxygenation response and survival in preterm infants receiving rescue inhaled nitric oxide therapy for hypoxemia from pulmonary hypertension: A systematic review and meta-analysis. Early Hum Dev. 2023 Sep;184:105841.

9. Bobrovitz N, Heneghan C, Onakpoya I, Fletcher B, Collins D, Tompson A, et al. Medications that reduce emergency hospital admissions: an overview of systematic reviews and prioritisation of treatments. BMC Med. 2018 Dec;16(1):115.

10. Mascarenhas D, Weisz D, Jasani B, Persad N, Main E. Premedication for rapid sequence intubation in neonates - a network meta-analysis. PROSPERO 2022 CRD42022384259 [Internet]. PROSPERO. Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022384259

11. Cochran WG. Sampling techniques, 3rd edition. John Wiley; 2002.

12. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. J Med Internet Res. 2024 Jan 12;26:e48996.

13. Tran VT, Gartlehner G, Yaacoub S, Boutron I, Schwingshackl L, Stadelmaier J, et al. Sensitivity and Specificity of Using GPT-3.5 Turbo Models for Title and Abstract Screening in Systematic Reviews and Meta-analyses. Ann Intern Med. 2024 Jun;177(6):791–9.

14. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine [Internet]. arXiv; 2023 [cited 2024 Jun 1]. Available from: https://arxiv.org/abs/2311.16452

15. Chai KEK, Lines RLJ, Gucciardi DF, Ng L. Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews. Syst Rev. 2021 Dec;10(1):93.

16. Perlman-Arrow S, Loo N, Bobrovitz N, Yan T, Arora RK. A real-world evaluation of the implementation of NLP technology in abstract screening of a systematic review. Res Synth Methods. 2023 Jul;14(4):608–21.

17. Nussbaumer-Streit B, Ellen M, Klerings I, Sfetcu R, Riva N, Mahmić-Kaknjo M, et al. Resource use during systematic review production varies widely: a scoping review. J Clin Epidemiol. 2021 Nov;139:287–96.

18. Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. Res Synth Methods. 2019 Sep;10(3):330–42.

19. Renze M, Guven E. The Effect of Sampling Temperature on Problem Solving in Large Language Models [Internet]. arXiv; 2024 [cited 2024 Aug 2]. Available from: https://arxiv.org/abs/2402.05201

20. Kojima T, Gu S (Shane), Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in neural information processing systems [Internet]. Curran Associates, Inc.; 2022. p. 22199–213. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c1 6f326-Paper-Conference.pdf

STARD Checklist

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	6
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	3
	4	Study objectives and hypotheses	2
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	2
Participants	**6**	Eligibility criteria	N/A
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	N/A
	8	Where and when potentially eligible participants were identified (setting, location and dates)	N/A
	9	Whether participants formed a consecutive, random or convenience series	N/A
Test methods	**10a**	Index test, in sufficient detail to allow replication	2
	10b	Reference standard, in sufficient detail to allow replication	4
	11	Rationale for choosing the reference standard (if alternatives exist)	N/A
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	N/A (binary)
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	N/A (binary)
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	4
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	4
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	6
	15	How indeterminate index test or reference standard results were handled	N/A (binary)
	16	How missing data on the index test and reference standard were handled	4
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	10
	18	Intended sample size and how it was determined	4
RESULTS			
Participants	**19**	Flow of participants, using a diagram	Figure 2
	20	Baseline demographic and clinical characteristics of participants	Figure 2, Table 1
	21a	Distribution of severity of disease in those with the target condition	Figure 2, Table 1
	21b	Distribution of alternative diagnoses in those without the target condition	Figure 2, Table 1
	22	Time interval and any clinical interventions between index test and reference standard	N/A

Test results	23	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	Table 3, 4
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	Figure 2, Table 1
	25	Any adverse events from performing the index test or the reference standard	N/A
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	12
	27	Implications for practice, including the intended use and clinical role of the index test	11
OTHER INFORMATION			
	28	Registration number and name of registry	N/A
	29	Where the full study protocol can be accessed	N/A
	30	Sources of funding and other support; role of funders	47