**­OpenAI’s GPT API models Can Function as a Highly Reliable Second Screener of Titles and Abstracts in Systematic Reviews**

**ABSTRACT**

Independent human double-screening of titles and abstracts is considered a pivotal step to ensure the quality of systematic reviews and meta-analyses. Yet, double screening is a costly as well as a time- and resource-intensive procedure that slows the review process, ultimately excluding many researchers from using it. To alleviate this issue and potentially increase the reliability of systematic reviews and meta-analyses, we evaluated the use of OpenAI’s GPT API models as an alternative second screener of titles and abstracts in large-scale systematic reviews. Overall, we found that GPT models perform similarly or even better than human screeners in terms of detecting relevant studies to be included. To support future reviewers, we develop a reproducible workflow and tentative guidelines for when reviewers can use GPT models for title and abstract screening. For this purpose, we present the R package AIscreenR.

GPT-4 models seem to work better FINDINGS

FURTHER DETAILS

To access the performance of ChatGPT, we develop benchmarks to compare the screening performance between humans and any given AI screener based on conflict rates estimates from 16 large-scale Campbell Systematic Reviews conducted over the last 10 years by the Danish VIVE Campbell group. In contrast to the typical conflict rate between human screeners, we find that ChatGPT can function, as a highly reliable second screener, which performs equally or better (i.e., fewer false excluded references) relative to human screeners.

**KEYWORDS:** *title and abstract screening, GPT API models, systematic review, screening benchmarks, AIscreenR*

**HIGHLIGHTS**

**What is already known**

* OpenAI’s GPT API models have been shown to work as a second screener of titles and abstracts within clinical and software literature.
* Automating screening tools can ease the burden of title and abstract screening

**What is new**

* We show that OpenAI’s GPT API models can function as a highly reliable second screener in social science reviews with better recalls than presented in previous evaluations.
* We develop general benchmarks to compare the performance of AI screening tools with human screening.
* We provide general guidelines for how and when GPT models safely can used
* We present and validate the R package AIscreenR to ensure standardized conduct of title and abstract screening with OpenAI’s GPT models.

**Potential impact for Research Synthesis Methods readers**

* Changing the workflow for double screening of title and abstract screening in systematic reviews
* Substantial reduction in human labor in systematic reviews
* Standardizing screening with GPT API models

**INTRODUCTION**

General intro

An all-important step to ensure the quality of systematic reviews and meta-analyses herein involves detecting all relevant references related to the literature under review (Polanin et al., 2019). Usually, this involves independent human double screening of all references detected in relevant databases and literature with two human screeners. This procedure has shown pivotally since less experienced single screeners tend to miss around 13% of relevant studies (with 3% for experienced screeners), mostly changing the main review findings (Waffenschmidt et al., 2019). Yet, double-screening is a costly and resource-intensive procedure which ultimately excludes many researchers from using it. An alternative to human double-screening is to use automated tools to act as the second screener (Gartlehner et al., 2019; Van De Schoot et al., 2021). Previous evaluations of existing tools find that most automated tools fail to reliably act as/imitating a human second screener. Meanwhile, it is still less known how well or if the newly developed large-language models (LLMs), such as ChatGPT, can work as a reliable second screener, especially within social science reviews.

Human in duplicate considered to be gold standard (Higgins et al., 2019; Wang et al., 2020). MOVE UP

We still need to know if it can work for social science reviews, there is a lack of guidelines for how to use GPT plus we software has not yet been develop to scale up the use of these models for title and abstract screening.

“*Deployment and user acceptance: requires (a) functioning tech (b) proof that it is functioning appropriately (c) the tech embodied in usable products (d) agreed guidelines for appropriate use (e) training (f) ongoing support.*” ([Campbell Collaboration](https://www.campbellcollaboration.org/news-and-events/news/stepping-up-evidence-synthesis.html))

We focus on proving (b) and developing as well as providing software and user guidelines to fulfill (c) and (d).

(Campos et al., 2023; Hou & Tipton, 2024)

*Previous research*

Syriani et al. (Syriani et al., 2023) test the performance relative to other machine-learning models.

*What we do differently*

* Use newest models with function calling. We are the first to present results for the GPT-4 model.

”*Function calling allows developers to more reliably get structured data back from the model.*” (<https://openai.com/blog/function-calling-and-other-api-updates>)

* Instead of comparing GPT to other machine learning models, we develop benchmark for comparing human and AI performance.
* Develop new software (AIscreenR) to standardize the title and abstract screening with GPT.
* Multi-core process to increase the time used on screening.
* Draw on function calling an incorporates uncertain decisions. “Function calling allows developers to more reliably get structured data back from the model” (<https://openai.com/blog/function-calling-and-other-api-updates>)

*Metrics we use to evaluate the performance of the gpt models*

All metrics presented below were chosen based on the recommendations made by Syriani et al. (2023). The two main metrics we used to evaluate the performance of the GPT API models were the recall and specificity metrics since these are intuitive to understand and interpret and are not sensitive to imbalanced data (i.e., data with a large discrepancy between inclusion and exclusion references). The recall “represents the proportion of relevant records being correctly classified” (Hou & Tipton, 2024), and can be written as

where (true positive) represents all the studies that are correctly included, and (false negative) is the number of studies falsely excluded. By contrast, specificity “measures the ability to exclude all references that should be excluded” (Syriani et al., 2023), given by

where (true negative) represents all the studies that are correctly excluded, and (false positive) is the number of studies falsely included. The recall metric can be considered the most important metric since it can seriously bias a review if the screener excludes references that should have been included. [FIND FURTER REASONS IN HOU & TIPTON] Whereas, a low specificity “just” means that reviewers must re-examine a larger share of the reference. This goes without saying that reviewers should accept low specificity rates. We will come back to that in the following sections.

We applied to overall assessment metric deduced from the above measure: mention imbalanced data

In our simulation, the , , , and conditions are determined by comparing the GPT decision with the final decision made by a minimum of two independent human screeners. For benchmark development, the conditions are determined by comparing the single screener decision with the final decision agreed upon between a minimum of two human screeners.

Mention how to calculate variance and confidence intervals. Viectbauer and Research synthesis methods. (Röver & Friede, 2022; Schwarzer et al., 2019)

Mention the nMCC model and formula

*Human performance vs. AI performance*

To grasp a better understanding of the AI performance.

We think it is more fair to compare the performance of the GPT models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source**  **Authors** | **Short title** | **Double-screened refs** | **Ass.** | **Aut.** |
| *Campbell review* |  |  |  |  |
| Bøg et al. (2018) | Deployment of personnel to military operations | 2899 | 2 | - |
| Bondebjerg et al. (2023) | The effects of small class sizes on students’ academic achievement, socioemotional development and well‐being in special education | 11860 | 4 | 2 |
| Dalgaard, Bondebjerg, Klokker et al. (2022) | Adult/child ratio and group size in early childhood education or care to promote the development of children aged 0–5 years | 3667 | 4 | 2 |
| Dalgaard, Bondebjerg, Viinholt et al. (2022) | The effects of inclusion on academic achievement, socioemotional development and wellbeing of children with special educational needs | 14491 | 5 | 2 |
| Dalgaard, Filges et al. (2022) | Parenting interventions to support parent/child attachment and psychosocial adjustment in foster and adoptive parents and children | 13106 | 3 | 2 |
| Dalgaard, Jensen et al. (2022) | PROTOCOL: Group‐based community interventions to support the social reintegration of marginalised adults with mental illness | 17614 | 4 | 3 |
| Dietrichson et al. (2020, 2021) | Targeted school-based interventions for improving reading and mathematics for students with or at risk of academic difficulties in Grades K-6 [plus 7-12] | 15273 | 6 | 1 |
| Filges, Dalgaard et al. (2022) | Outreach programs to improve life circumstances and prevent further adverse developmental trajectories of at-risk youth in OECD countries | 4890 | 4 | - |
| Filges, Dietrichson et al. (2022) | Service learning for improving academic success in students in grade K to 12 | 6269 | 4 | 1 |
| Filges, Montgomery, et al. (2015) | The Impact of Detention on the Health of Asylum Seekers | 10061 | 2 | - |
| Filges, Siren et al. (2020) | Voluntary work for the physical and mental health of older volunteers | 14919 | 2 | 0 |
| Filges, Smedslund et al. (2023) | PROTOCOL: The FRIENDS preventive programme for reducing anxiety symptoms in children and adolescents | 2745 | 1 | 1 |
| Filges, Sonne-Schmidt et al. (2018) | Small class sizes for improving student achievement in primary and secondary schools | 7802 | 5 | 1 |
| Filges, Torgerson, et al. (2019) | Effectiveness of continuing professional development training of welfare professionals on outcomes for children and young people | 5147 | 1 | 4 |
| Filges, Verner et al. (2023) | PROTOCOL: Participation in organised sport to improve and prevent adverse developmental trajectories of at-risk youth | 7796 | 2 | 1 |
| *NIPH review* |  |  |  |  |
| Ames et al. (2024) | Acceptability, values, and preferences of older people for chronic low back pain management | 6644 | - | 2 |
| Evensen et al. (2023) | Sutur av degenerative rotatorcuff-rupturer [Rotator cuff repair for degenerative rotator cuff tears] | 425 | - | 4 |
| Jardim et al. (2021) | Effekten av antipsykotika ved førstegangspsykose [The effect of antipsychotics on first episode psychosis] | 3924 | - | 3 |
| Johansen et al. (2022) | Samværs-og bostedsordninger etter samlivsbrudd [Custody and living arrangements after parents separate] | 1525 | - | 4 |
| Meneses Echavez et al. (2022) | Psykologisk debriefing for helsepersonell involvert i uønskede pasienthendelser [Psychological debriefing for healthcare professionals involved in adverse events] | 5452 | - | 3 |

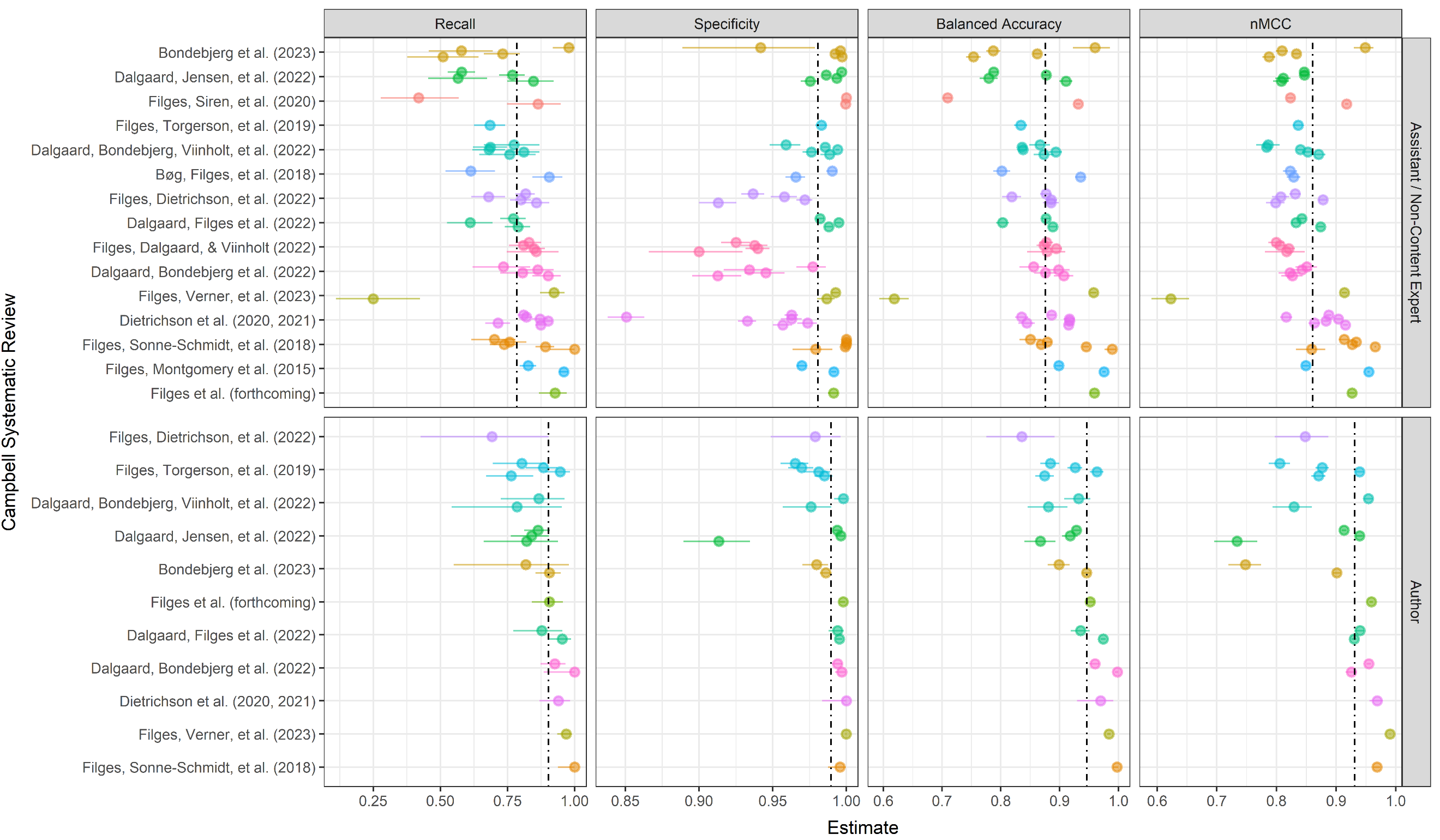


FIGURE 1. Performance measures within Campbell Systematic Reviews across assistants vs. authors. Dashed line indicate the average estimated via the SCE+ model.



FIGURE 2. Researcher-researcher screening performance measures within NIPH Systematic Reviews. Dashed line indicate the average estimated via the CHE model.

Mention the authority and deeper content knowledge of the main author which might cause the recall to increase when review author screen with student assistants. Therefore to compare screenings with more equal relations, we analyze data from sixe systematic reviews conducted by the Norwegian Institue of Public Health (NIPH).

Imbalance not a problem with GPT models cf. FRIENDS

*Simulation data*

FRIENDS and FTT

*Prompt engineering*

[insert prompt example]

*The simulation results*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Review**  **Model** | **Reps** | **Recall (%)**  **[TP/]a** | **Specificity (%)**  **[TN/]b** | **Raw aggrement (%)**  **[(TP + TN)/]c** | **bAcc (%)** |
| *FFT* |  |  |  |  |  |
| GPT-3.5-turbo-0613  (incl. prop = .5) | 10 | 71  (49/69) | 95.6  (3888/4066) | 95.2  (3937/4135) | 83.3 |
| GPT-3.5-turbo-0613  (incl. prop = .3) | 10 | 81.2  (56/69) | 93.7  (3809/4066) | 93.5  (3865/4135) | 87.4 |
| GPT-4-0613 | 1 | 89.9  (62/69) | 93.7  (3810/4066) | 93.6  (3872/4135) | 91.8 |
| *FRIENDS* |  |  |  |  |  |
| GPT-3.5-turbo-0613  (incl. prop = .5) | 10 | 96.9  (62/64) | 76.5  (1930/2511) | 77.1  (1992/2575) | 86.7 |
| GPT-3.5-turbo-0613  (incl. prop = .7) | 10 | 95.3  (61/64) | 89.8  (2256/2511) | 90.0  (2317/2575) | 92.6 |
| GPT-4-0613 | 1 | 98.4  (63/64) | 97.4  (2455/2511) | 97.4  (2518/2585) | 97.9 |

*a*: is all included references, *b*: is all excluded references, *c*: is the total number of references

Concise text more important than information-dense prompt.

The GPT-3.5-turbo model is sensitive to the number of times a study is included across 10 iterations.

In contrast with priority screening methods(Hou & Tipton, 2024), the gpt models do have the potential to find more than 95% of the relevant study cf. FRIENDS.

A side-goal of this simulation was also to validate the performance of the AIscreenR software. Especially the use of function calling.

*Tentative guidelines*

*Workflow and short package presentation*

*Deficits of using GPT API models*

* Black box (but this does not only count for GPT this is often true for human screening as well)
* Function tech? We have no control over the existence of OpenAI

*Future research*

* The use of hirarchical prompting in complex reviews

*Discussion*

* *Talk about interface here – cannot replicate the results on the ChatGPT interface*
* *Reviewers should not consider screening prioritization methods and GPT screening as two incommensurable methods. Instead, the strength from both should ideally be combined.*
* *Forces review times to make very narrow searches due to lack of ressources to conduct the title and abstract screening rigourously (Guo find in ICloud)*
* *We believe that the GPT-4 models will perform even better when fed with abstracts following a rigorous structure as in medicine.*

**ACKNOWLEDGEMENT**

Thanks to Trine Filges, Jens Dietrichson, Tiril Borge, Heather Melanie R. Ames, and Christopher James Rose for valuable comments and sharing of screen data. Also thanks to Sofie Elgaard Lisager Jensen, Johan Klejs, and Frederikke Lykke Witthöft Schytt for testing the AIscreenR software and for valuable inputs to the workflow.

**CONFLICT OF INTEREST STATEMENT**

The authors declare no conflict of interest.

**DATA AVAILABILITY STATEMENT**

Insert OSF or GitHub page

**REFERENCES**

Ames, H., Hestevik, C. H., & Briggs, A. M. (2024). Acceptability, values, and preferences of older people for chronic low back pain management; a qualitative evidence synthesis. *BMC Geriatrics*, *24*(1), 1–22. https://doi.org/10.1186/s12877-023-04608-4

Bøg, M., Filges, T., & Jørgensen, A. M. K. (2018). Deployment of personnel to military operations: impact on mental health and social functioning. *Campbell Systematic Reviews*, *14*(1), 1–127. https://doi.org/https://doi.org/10.4073/csr.2018.6

Bondebjerg, A., Dalgaard, N. T., Filges, T., & Viinholt, B. C. A. (2023). The effects of small class sizes on students’ academic achievement, socioemotional development and well‐being in special education: A systematic review. *Campbell Systematic Reviews*, *19*(3), e1345.

Campos, D. G., Fütterer, T., Gfrörer, T., Lavelle-Hill, R. E., Murayama, K., König, L., Hecht, M., Zitzmann, S., & Scherer, R. (2023). *Screening Smarter, Not Harder: A Comparative Analysis of Machine Learning Screening Algorithms and Heuristic Stopping Criteria for Systematic Reviews in Educational Research*.

Dalgaard, N. T., Bondebjerg, A., Klokker, R., Viinholt, B. C. A., & Dietrichson, J. (2022). Adult/child ratio and group size in early childhood education or care to promote the development of children aged 0–5 years: A systematic review. *Campbell Systematic Reviews*, *18*(2), e1239. https://doi.org/https://doi.org/10.1002/cl2.1239

Dalgaard, N. T., Bondebjerg, A., Viinholt, B. C. A., & Filges, T. (2022). The effects of inclusion on academic achievement, socioemotional development and wellbeing of children with special educational needs. *Campbell Systematic Reviews*, *18*(4), e1291. https://doi.org/https://doi.org/10.1002/cl2.1291

Dalgaard, N. T., Filges, T., Viinholt, B. C. A., & Pontoppidan, M. (2022). Parenting interventions to support parent/child attachment and psychosocial adjustment in foster and adoptive parents and children: A systematic review. *Campbell Systematic Reviews*, *18*(1), e1209. https://doi.org/https://doi.org/10.1002/cl2.1209

Dalgaard, N. T., Flensborg Jensen, M. C., Bengtsen, E., Krassel, K. F., & Vembye, M. H. (2022). PROTOCOL: Group‐based community interventions to support the social reintegration of marginalised adults with mental illness. *Campbell Systematic Reviews*, *18*(3), e1254. https://doi.org/10.1002/cl2.1254

Dietrichson, J., Filges, T., Klokker, R. H., Viinholt, B. C. A., Bøg, M., & Jensen, U. H. (2020). Targeted school-based interventions for improving reading and mathematics for students with, or at risk of, academic difficulties in Grades 7–12: A systematic review. *Campbell Systematic Reviews*, *16*(2), e1081. https://doi.org/10.1002/cl2.1081

Dietrichson, J., Filges, T., Seerup, J. K., Klokker, R. H., Viinholt, B. C. A., Bøg, M., & Eiberg, M. (2021). Targeted school-based interventions for improving reading and mathematics for students with or at risk of academic difficulties in Grades K-6: A systematic review. *Campbell Systematic Reviews*, *17*(2), e1152. https://doi.org/10.1002/cl2.1152

Evensen, L. H., Kleven, L., Dahm, K. T., Hafstad, E. V., Holte, H. H., Robberstad, B., & Risstad, H. (2023). *Sutur av degenerative rotatorcuff-rupturer: en fullstendig metodevurdering [Rotator cuff repair for degenerative rotator cuff tears: a health technology assessment].* https://www.fhi.no/publ/2023/sutur-av-degenerative-rotatorcuff-rupturer/

Filges, T., Dalgaard, N. T., & Viinholt, B. C. A. (2022). Outreach programs to improve life circumstances and prevent further adverse developmental trajectories of at-risk youth in OECD countries: A systematic review. *Campbell Systematic Reviews*, *18*(4), e1282. https://doi.org/https://doi.org/10.1002/cl2.1282

Filges, T., Dietrichson, J., Viinholt, B. C. A., & Dalgaard, N. T. (2022). Service learning for improving academic success in students in grade K to 12: A systematic review. *Campbell Systematic Reviews*, *18*(1), e1210. https://doi.org/https://doi.org/10.1002/cl2.1210

Filges, T., Montgomery, E., Kastrup, M., & Jørgensen, A.-M. K. (2015). The Impact of Detention on the Health of Asylum Seekers: A Systematic Review. *Campbell Systematic Reviews*, *11*(1), 1–104. https://doi.org/https://doi.org/10.4073/csr.2015.13

Filges, T., Siren, A., Fridberg, T., & Nielsen, B. C. V. (2020). Voluntary work for the physical and mental health of older volunteers: A systematic review. *Campbell Systematic Reviews*, *16*(4), e1124. https://doi.org/https://doi.org/10.1002/cl2.1124

Filges, T., Smedslund, G., Eriksen, T., & Birkefoss, K. (2023). PROTOCOL: The FRIENDS preventive programme for reducing anxiety symptoms in children and adolescents: A systematic review. *Campbell Systematic Reviews*, *19*(4), e1374. https://doi.org/https://doi.org/10.1002/cl2.1374

Filges, T., Sonne‐Schmidt, C. S., & Nielsen, B. C. V. (2018). Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *Campbell Systematic Reviews*, *14*(1), 1–107. https://doi.org/10.4073/csr.2018.10

Filges, T., Torgerson, C., Gascoine, L., Dietrichson, J., Nielsen, C., & Viinholt, B. A. (2019). Effectiveness of continuing professional development training of welfare professionals on outcomes for children and young people: A systematic review. *Campbell Systematic Reviews*, *15*(4), e1060. https://doi.org/https://doi.org/10.1002/cl2.1060

Filges, T., Verner, M., Ladekjær, E., & Bengtsen, E. (2023). PROTOCOL: Participation in organised sport to improve and prevent adverse developmental trajectories of at-risk youth: A systematic review. *Campbell Systematic Reviews*, *19*(2), e1321. https://doi.org/https://doi.org/10.1002/cl2.1321

Gartlehner, G., Wagner, G., Lux, L., Affengruber, L., Dobrescu, A., Kaminski-Hartenthaler, A., & Viswanathan, M. (2019). Assessing the accuracy of machine-assisted abstract screening with DistillerAI: a user study. *Systematic Reviews*, *8*(1), 277. https://doi.org/10.1186/s13643-019-1221-3

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M. S., Li, T., Page, M., & Welch, V. (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). Wiley Online Library. https://doi.org/10.1002/9781119536604

Hou, Z., & Tipton, E. (2024). Enhancing recall in automated record screening: A resampling algorithm. *Research Synthesis Methods*, *n/a*(n/a). https://doi.org/https://doi.org/10.1002/jrsm.1690

Jardim, P. S. J., Borge, T. C., & Johansen, T. B. (2021). *Effekten av antipsykotika ved førstegangspsykose: en systematisk oversikt [The effect of antipsychotics on first episode psychosis]*. https://fhi.no/publ/2021/effekten-av-antipsykotika-ved-forstegangspsykose/

Johansen, T. B., Nøkleby, H., Langøien, L. J., & Borge, T. C. (2022). *Samværs-og bostedsordninger etter samlivsbrudd: betydninger for barn og unge: en systematisk oversikt [Custody and living arrangements after parents separate: implications for children and adolescents: a systematic review]*. https://www.fhi.no/publ/2022/samvars--og-bostedsordninger-etter-samlivsbrudd-betydninger-for-barn-og-ung/

Meneses Echavez, J. F., Borge, T. C., Nygård, H. T., Gaustad, J.-V., & Hval, G. (2022). *Psykologisk debriefing for helsepersonell involvert i uønskede pasienthendelser: en systematisk oversikt [Psychological debriefing for healthcare professionals involved in adverse events: a systematic review]*. https://www.fhi.no/publ/2022/psykologisk-debriefing-for-helsepersonell-involvert-i-uonskede-pasienthende/

Polanin, J. R., Pigott, T. D., Espelage, D. L., & Grotpeter, J. K. (2019). Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Research Synthesis Methods*, *10*(3), 330–342. https://doi.org/https://doi.org/10.1002/jrsm.1354

Röver, C., & Friede, T. (2022). Double arcsine transform not appropriate for meta-analysis. *Research Synthesis Methods*, *13*(5), 645–648. https://doi.org/https://doi.org/10.1002/jrsm.1591

Schwarzer, G., Chemaitelly, H., Abu-Raddad, L. J., & Rücker, G. (2019). Seriously misleading results using inverse of Freeman-Tukey double arcsine transformation in meta-analysis of single proportions. *Research Synthesis Methods*, *10*(3), 476–483. https://doi.org/https://doi.org/10.1002/jrsm.1348

Syriani, E., David, I., & Kumar, G. (2023). Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews. *ArXiv Preprint ArXiv:2307.06464*.

Van De Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., & Ferdinands, G. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2), 125–133.

Waffenschmidt, S., Knelangen, M., Sieben, W., Bühn, S., & Pieper, D. (2019). Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC Medical Research Methodology*, *19*(1), 132. https://doi.org/10.1186/s12874-019-0782-0

Wang, Z., Nayfeh, T., Tetzlaff, J., O’Blenis, P., & Murad, M. H. (2020). Error rates of human reviewers during abstract screening in systematic reviews. *PloS One*, *15*(1), e0227742.