A protocol to evaluate unsupervised text clustering to screen and categorize studies in systematic reviews

Ashley Elizabeth Muller (aemu@fhi.no)
Norwegian Institute of Public Health https://orcid.org/0000-0001-7819-6697
Heather Melanie R Ames
Tiril C Borge
Christine Hillestad Hestevik
Jose Francisco Meneses-Echavez
Christopher James Rose

Method Article

Keywords: title and abstract, artificial intelligence, evidence synthesis, software tools, information management, methodology, trust, user experience

Posted Date: May 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1644531/v1

Abstract

Background: The evidence base is steadily increasing regarding the time savings of using machine learning (ML) within evidence synthesis, particularly supervised methods such as classification. Unsupervised methods such as clustering have been less explored. Yet clustering – a method to uncover groups of similar data, from a large and heterogenous dataset – may be a particularly relevant tool within the study identification and data extraction processes of reviews, during which researchers must often read thousands of studies to identify the handful relevant to their review. This protocol is for a mixed methods evaluation of clustering implemented within systematic reviews.

Research questions: We will answer the following questions: 1) Can clustering while screening at title/abstract level provide useful automatic categorizations of studies that help reviewers extra data to categorize studies? 2) What are reviewers' perceptions of and experiences with clustering in relation to acceptability and feasibility in future reviews?

Methods: We will identify one or two systematic reviews as models to answer these research questions. Reviewers will use the clustering algorithm Lingo3G built into EPPI Reviewer during the screening process. The main outcome to answer research question 1 is content validity, as an indicator of usefulness, defined by similarities between automatic clusters and reviewer-generated categories. We will display multi-class confusion matrices and use these to estimate precision, recall, and F1-score for each useful cluster. Research question 2 will be answered qualitatively, through semi-structured interviews with participating reviewers before and after having been trained on clustering and implementing it in the model review(s), followed by a focus group presenting results and exploring reviewer trust.

Conclusion: The planned evaluation will provide important information as to the feasibility of an unsupervised ML method, clustering, in bridging and aligning two currently linear steps in evidence synthesis. We invite interested evidence synthesis organizations to follow this protocol and subsequently share data with us, or to tailor it to produce a more relevant evaluation.

1. Background

Evidence synthesis provides the foundation for evidence-based medicine and policy, yet traditional approaches cannot meet policymakers' needs for evidence (Bastian, Glasziou, & Chalmers, 2010; Elliott et al., 2017). Westgate and colleagues have called this the "synthesis gap" (Westgate et al., 2018). The COVID-19 pandemic appears to have led to a step-change in the adoption of machine learning (ML) within evidence synthesis, perhaps particularly of health research (Blaizot et al., 2022). As a rough estimate of increasing interest and use, Figure 1 below displays the increase in review protocols registered on PROSPERO with two relevant terms, registered in the year before during the pandemic, and again in the second year of the pandemic.

Figure 1 Reviews registered on PROSPERO before and after the COVID-19 pandemic began, with various ML-related terms

Figure 1 Legend: The amount of reviews registered on PROSPERO with various machine learning-related terms has increased by at least a factor of four from the year before COVID-19, to the second year of the pandemic. 2017 amounts were the reviews added to PROSPERO 30 Nov 2016 – 30 Nov 2017; 2019 amounts were the reviews added 30 Nov 2018 – 30 Nov 2019; and 2021 amounts corresponded to 30 Nov 2020 – 30 Nov 2021.

ML was certainly not underdeveloped before the pandemic, and numerous evidence synthesis research groups included programmers and ML specialists (see for example (Beller et al., 2018; Haddaway et al., 2020; Haddaway et al., 2019; Haddaway & Westgate, 2020; Kohl et al., 2018; O'Connor et al., 2019; O'Mara-Eves, Thomas, McNaught, Miwa, & Ananiadou, 2015; J. Thomas et al., 2017; J Thomas & Stansfield, 2018)). To aide implementation within the field, they also published proof of concept articles, case studies of overall approaches and off-the-shelf systems, and software evaluations (Harrison, Griffin, Kuhn, & Usher-Smith, 2020; Marshall, Johnson, Wang, Rajasekaran, & Wallace, 2020; Marshall, Kuiper, Banner, & Wallace, 2017; Marshall, Kuiper, & Wallace, 2016; Stansfield, Thomas, & Kavanagh, 2013) as well as practical guidelines (Haddaway et al., 2020; Marshall & Wallace, 2019; Polanin, Pigott, Espelage, & Grotpeter, 2019).

During and as a response to the pandemic, these techniques have been further developed and rapidly scaled up to tackle the deluge of COVID-19 research (Agai, 2020; Rada et al., 2020; I Shemilt et al., 2021; Shemilt, Noel-Storr, Thomas, Featherstone, & Mavergames, 2021). Evidence synthesis organizations, research groups, dedicated journals, and influential guideline-setters have begun providing guidance and best practices for the use of ML (Haddaway et al., 2020; Hamel et al., 2021; Khalil, Tamara, Rada, & Akl, 2021; Page et al., 2021). Other groups and dedicated organizations have begun using ML for the first time.

Use of clustering within evidence synthesis

Clustering is a well-known form of unsupervised machine learning, the goal of which is to “to reveal subgroups within heterogeneous data such that each individual cluster has greater homogeneity than the whole” (Alashwal, El Halaby, Crouse, Abdalla, & Moustafa, 2019, p. 2). When applied to a set of studies (here, to the text of titles and abstracts), each study is assigned to one or more automatically identified clusters such that any two studies within the same cluster are similar in some way, and studies in different clusters are dissimilar in some way. In other words, clusters help researchers to identify similar studies, after which they must decide whether that similarity is meaningful or useful to them.

Depending on the algorithm used, researchers may be able to rely on the cluster label to help understand a cluster's similarity characteristic; alternatively, it may be necessary to read the titles or abstracts to infer why the studies were assigned to a given cluster. The more manual inspection required, the more time-intensive this process is, a well-known drawback to unsupervised ML methods (Morichetta, 2019).

Berrang-Ford et al. (Berrang-Ford et al., 2021) recently used a related form of unsupervised ML, topic modelling, to systematically map existing literature on climate change and health. They compared two algorithms and multiple parameters to perform topic modelling, and proceeded with non-negative matrix factorization and 70 topics. This case study is impressive and allowed them to map and categorize more than 15,000 studies published in a six-year time span – but the methods require advanced ML skills, which is not a realistic expectation for a typical systematic reviewer. Less encouragingly, Hartmann et al. (Hartmann, Wuijts, van der Hoek, & de Roda Husman, 2019) evaluated the clustering software Adjutant in R and concluded that it was not a feasible approach to fully automate study selection in a review of emerging aquatic contaminates. Only 53% of the nearly 13,000 retrieved studies were clustered with Adjutant, and only 28% of an *a priori* selection of 14 relevant articles. As with Berrang-Ford et al., Hartmann's impressive methods also require advanced ML skills.

The ML implementation team and its evaluations

The Cluster for Reviews and Health Technology Assessments at the Norwegian Institute of Public Health is a new adopter of ML. As a direct result of the need to synthesize COVID-19 research for national health and welfare authorities at a dramatically faster pace, the cluster's 50 systematic reviewers and librarians went from using ML in no studies in 2019, to using ML in approximately half of all review products in 2020 (Muller, Ames, et al., 2021b).

This shift was evidence-based, as early explorations with COVID-19 reviews (Himmels, Borge, Brurberg, & Gravningen, 2021; Himmels, Gomez Castaneda, Brurberg, & Gravningen, 2021; Rost, Slaughter, Nytro, Muller, & Vist, 2021) and subsequently other social welfare and health topics (Hestevik, Muller, & Forsetlund, 2021; Muller, Ames, Jardim, & Rose, 2021) demonstrated time savings of both various ML functions. Since late 2020, the ML implementation team coordinates implementation of ML functions, including building the capacity of reviewers to independently use, interpret, and explain different ML functions (Muller, Ames, et al., 2021a). This team is also tasked with the continuous identification, evaluation, and implementation of ML functions and applications that can aid the production of evidence synthesis products. Finally, the team also coordinates changes communication around ML.

The ML implementation team supports the use of the clustering algorithm Lingo3G among systematic reviewers who do not have advanced ML skills. A benefit of this clustering algorithm is that it is more *description-centric* than *data-centric*. Clusters are automatically assigned descriptive labels (rather than numbers or other non-semantic identifiers), and users can request that labels are constructed using fewer or more words, which can help researchers understand the clustering (Osiński, Stefanowski, & Weiss,

2004; Osiński & Weiss, 2005). Another benefit is simply its availability; it is integrated into our systematic review software of choice, EPPI Reviewer (J Thomas et al., 2020).

One of the first explorations of the ML team was to evaluate the content validity of a single-hierarchy clustering set in Muller et al. (Muller, Ames, Jardim, et al., 2021). Briefly, we found excellent agreement in categorizations between a Lingo3G, an independent human, and a human checking the clustering algorithm (non-blinded), assessing all against the same gold standard. To quantitatively compare human and computer performance, we computed precision and recall. Clustering had similar precision to both independent and non-blinded researchers (e.g., 88% vs 89%), but higher recall (e.g., 89% vs 84%). We estimated that the fully automated clustering procedure reduced time use by 71% compared to the fully manual procedure, and the semi-automated procedure reduced time use by 34%.

We are not aware of any published explorations of clustering among reviewers who are not ML specialists aside from an early case study by Stansfield and colleagues (Stansfield et al., 2013). As in Muller et al., Stansfield and colleagues examined how useful clusters were to answer pre-defined research questions within two reviews. Clusters were relevant to seven of their eight research questions, while we found that clusters were relevant to only two of our four broad thematic categories.

Need for further evaluation of clustering

Our content validity evaluation suggests that when automatically generated clusters are deemed useful, we can probably trust the distribution of studies in these clusters. Because clustering is unsupervised, researchers cannot define a categorization system *a priori*, so not all automatically generated clusters will be useful. Previous evaluations show different levels of "usefulness" (e.g., 88% in Stansfield et al. and 50% in Muller et al.) There is also no guarantee that all desired useful categories will be generated automatically. In other words, it will not be possible to use automatic clustering in all projects, and it is difficult or impossible to know this before trying to use the method. This has implications for project planning: while use of automated tools may be expected to save time on average, it is likely to increase the variance in time use between projects compared to current practice, which means it will be harder to predict how much time a specific project will need.

We therefore need to know whether clustering can help to combine screening and data-mapping phases by automatically identifying obviously relevant or irrelevant studies, uncovering previously unidentified study similarities, and grouping studies according to organic but useful categories. Weißer et al. (Weisser, Sassmannshausen, Ohrndorf, Burggraf, & Wagner, 2020) published a methods paper that argues for using clustering to bridge screening and categorization: "By using the method presented in this work, the task of the reviewer is not to read abstracts and filter articles, but to exclude clusters of articles respectively to integrate clusters of articles of low/high relevance for the topic of interest" (p. 9). Hamel et al. (Hamel et al., 2021) have recently published best practices for screening with ML, one of which is to consider "optimizing" the team by conducting data extraction in tandem with screening. Their

recommendation was related to stopping or deprioritizing title/abstract screening once most relevant studies have been identified. Our evaluation will go one step further, by exploring whether clustering can help uncover content similarity and differences between studies during screening, and if these similarities and differences are themselves useable in the data extraction phase.

Separate from empirical usefulness and potential to improve workflows, is whether researchers accept this usefulness and are willing to use clustering in future reviews. In an unpublished mixed methods study we conducted in 2021, a ML system was estimated to perform equal to or better than humans in assessing risk of bias of randomized controlled trials. Yet, even after being presented with empirical evidence of the potential superiority of ML, the reviewers who participated in this study said they trusted human assessments more. The measurement of human acceptability – or lack thereof – is thus as important to a ML evaluation as the measurement of performance.

Objective

This protocol is for a mixed methods evaluation of clustering within systematic reviews. We will identify one or two systematic reviews, and within these reviews, explore whether clustering can bridge two distinct steps of a review (study identification and data extraction), as well as whether participating researchers accept this process. We invite interested evidence synthesis organizations to follow this protocol and share data, or to tailor it to produce an evaluation that addresses their unique research questions.

2. Methodology

This protocol is a plan to answer the following research questions:

Research question 1: Can clustering while screening at title/abstract level provide useful categorizations of studies that help reviewers extract data to categorize or map studies?

Research question 2. What are reviewers' perceptions of and experiences with clustering in relation to acceptability and feasibility in future reviews?

Eligibility requirements

One or more systematic reviews can be used as models to answer these research questions, with the following characteristics:

Review type: Any review that relies on titles and abstracts for data mapping or that is willing to start this process at the title and abstract level, such as a scoping review, rapid review, or evidence and gap map.

Other review characteristics: The review must have a research question(s) that will be answered through categorization or mapping, or an aim to uncover topics that can be further refined or changed. A review with strictly predefined categories is not appropriate.

Review team characteristics: Participating reviewers must be willing to combine two review phases (screening and data mapping/categorization). They must be willing to engage with this ML function, experiment with various parameters, and have manageable expectations of the iterative nature of clustering, rather than expecting clustering to perform according to a predefined structure.

Support from ML team: The ML team anticipates providing the entire review team with two to four hours of training and support. As per ML team procedures, support will be provided by one designated ML contact.

Procedures

This protocol can be followed from the screening phase of the review, after references have been retrieved from systematic searches and de-duplicated. The review team will receive training from the ML team on the use of clustering. As per the ML team's implementation procedures, such training includes a 45-60 minute "conceptual" session of a ML function, here describing in plain language what clustering is and how to interpret clusters, followed by a 45-60 minute "technical" session, which teaches how to practically set up, run, and document the output of clustering, as well as how to quality-control and check for red flags that could indicate something has gone wrong.

The review lead will apply the clustering algorithm Lingo3G within the EPPI Reviewer software as often as deemed necessary, and at least at the beginning of screening, upon receiving database-identified studies from librarians.

We anticipate that the review team uses clusters to simultaneously conduct what is typically a linear process, namely screening and data extraction. The review team will track all clustering structures which they have deemed useful as codesets in EPPI Reviewer. Modifiable parameters include minimum and maximum cluster size, maximum hierarchy depth, minimum label length, and single word label weight. The review lead will describe the parameters of all retained cluster structures. Within retrained clustering structures, the review lead will re-name clusters and discard clusters as appropriate and record the phase the clusters were used in, whether screening, data-mapping, or combined activities. The review lead will decide how to proceed with all clusters, e.g., machine-screen, prioritize for screening, prioritize for coding, auto-code or -categorize, etc.

This evaluation can be conducted within the normal production of a review. See Figure 2 for the steps of both this evaluation and the participating model review or reviews superimposed on each other. Quantitative analysis (see section 2.3) can begin once the review's included studies have received their

final coding/categorizations. Qualitative analysis (see section 2.4) can begin upon the completion of the first conversation and be completed after the group discussion has been transcribed.

Figure 2 The steps of this evaluation mapped to the phases of the model review(s)

Figure 2 legend: This evaluation can be followed within the timeline of the one or two model reviews that are participating.

Quantitative outcomes and analytic plan

Research question 1: Can clustering while screening at title/abstract (T/A) level provide useful categorizations of studies that help reviewers extract data to categorize or map studies?

The main outcome is content validity, as an indicator of usefulness. We will define content validity by similarities between automatic clusters and reviewer-generated clusters, similar to Stansfield et al.'s approach. Clusters used in data-mapping will be cross-tabulated against final coding schemes, such as the x- and y-axes in an evidence-and-gap map. We will display multi-class confusion matrices (see Figure 3 for an example) and use these to estimate precision (see Figure 4 or an example), recall, and F1-score for each useful cluster. AEM, TCB, or CJR will perform the analyses.

Figure 3 Sample confusion matrix

Figure legend: A sample confusion matrix showing three predicted classes and three true classes.

Figure 4 Precision values for the above confusion matrix

Figure legend: Precision values for the above sample confusion matrix with three predicted classes and three true classes.

Qualitative exploration and analytic plan

Research question 2: What are reviewers' perceptions of and experiences with clustering in relation to acceptability and feasibility?

We will explore acceptability and feasibility of this clustering as perceived by participating review team(s) using a repeated semi-structured interview with each reviewer, and a focus group discussion at the end of the project. We plan to record both the semi-structured interviews and the focus group discussion with the reviewers' consent and then transcribe for analysis.

Sampling

We plan to interview all reviewers who participate in the clustering training and who implement clustering as part of their systematic review (an estimated three reviewers per review).

Semi-structured interviews

HA and CHH will build a semi-structured interview guide drawing on two theoretical frameworks. The first is Roger's diffusion of technology framework, which was used successfully by Arno et al. (Arno, Elliott, Wallace, Turner, & Thomas, 2021) to explore health guideline makers' attitudes towards ML. The second is the "honeycomb framework" of user experience, which was originally developed by Peter Morville and adapted by Rosenbaum (S. Rosenbaum, 2010). This framework has been used and adapted based on findings from several similar studies exploring participants' experiences of technology designed to facilitate use of research evidence in health decision making (S. E. Rosenbaum, Glenton, & Cracknell, 2008; S. E. Rosenbaum, Glenton, Nylund, & Oxman, 2010; S. E. Rosenbaum et al., 2011). After the clustering training, HA or CHH will conduct individual semi-structured interviews with the reviewers involved to explore their perceptions of clustering, their understanding of how it works, and whether they feel it is a useful function. After the reviewers have implemented clustering within the actual review(s), a second interview will be conducted to examine and explore how their understanding and perceptions of clustering have evolved. This feedback will also be sent back to the ML team to improve the content of our training materials.

Focus group discussion

After analysis for research question 1 is complete, we will present results to participating reviewers, then hold a focus group discussion on their perceptions and interpretations of the findings. HA and CHH will develop a list of questions to guide the discussion based on the results of the analysis for research question 1 and the interviews. The focus group discussion will allow us to user check our findings and receive input on reaction to and understanding of the studies' findings.

Analysis

While we are developing the interview guide, we will also create the *a priori* framework for data analysis. HMRA and CHH will then conduct a best fit framework synthesis of the transcripts of the individual conversations and the group discussion (Booth & Carroll, 2015; Carroll, Booth, Leaviss, & Rick, 2013). A thematic analysis will be conducted on any data that does not fit into the *a priori* framework (Miles, Huberman, & Saldana, 2019; Silverman, 2013). The framework will then be adjusted and expanded to encompass the new themes and findings (Booth & Carroll, 2015).

Data administration

The audio recordings, taken after signed informed consent has been collected, will be stored in a private file on a password-protected institute computer and deleted once transcription has been finalized.

Dissemination and communication

We will disseminate results externally as a journal article. Nationally, we will present results to the Department of Environmental Health at NIPH; the Norwegian Scientific Community for Food and Environment; the ML/AI Network at NIPH led by AEM; and the Artificial Intelligence in Norwegian Health Services Network. Finally, we will share any English-language presentations with our institutional collaborators EPPI Centre and Julius Kühn Institut – Federal Research Centre for Cultivated Plants, and with the International Collaboration for the Automation of Systematic Reviews.

3. Conclusion And Call For Collaboration

Results from this study will help us make an important recommendation regarding the extent to which two phases in evidence synthesis – screening and data extraction – may be combined using assistance from unsupervised ML. Several ML functions consistently reduce time in various evidence synthesis phases (O'Mara-Eves et al., 2015; James Thomas et al., 2021). The potential to bridge and combine phases could result in further time savings.

Equally important will be knowledge gained about whether researchers accept this form of unsupervised ML, and why or why not. As we have found in a previous study evaluating a supervised ML system, satisfactory ML performance did not translate into acceptability among human researchers (manuscript submitted).

We also hope that this evaluation will inform recommendations to aid the explainability of unsupervised ML. Angelov et al. (Angelov, Soares, Jiang, Arnold, & Atkinson, 2021) describe general principles of explainable ML algorithms as indicating the opposite of a "'black box' nature", namely transparency and interpretability. As ML becomes an increasingly important tool to produce high-quality evidence synthesis rapidly, it is vital that researchers can explain these tools. Unexplainable ML, in contrast, may undermine trust in evidence syntheses (Arno et al., 2021).

Figure 5 Practical research agenda for unsupervised clustering

Figure legend: We suggest three directions for future research and practice around unsupervised clustering in evidence synthesis to explore. See also Muller et al. (Muller, Ames, Jardim, et al., 2021).

We invite research groups with reviews and review teams that meet evaluation requirements to contact us for collaboration. We intend for other groups to be able to adapt all of parts of this protocol to their own workflows. We are interested in pooling data across experiments with other research groups, as well as collaborating with groups who can take one or more parts of this protocol further, such as by expanding the qualitative analysis or delving further into measuring transparency, interpretability, and explainability of clustering. Finally, we invite critical feedback on this protocol.

We conclude by proposing a research agenda in Figure 5.

4. Declarations

Ethics approval and consent to participate

Ethical approval is not required as no personal or sensitive data will be collected, and the data will be anonymized. This evaluation will be conducted within production of a normal systematic review, lending no additional safety concerns to participants.

Availability of data and materials

Data produced through the implementation of the planned evaluation will be made publicly available.

Competing interests

The authors declare no competing interests.

Funding

The planned evaluation will be implemented as a normal activity of the ML team, which is funded by the Cluster for Reviews and Health Technology Assessment leadership.

Authors’ contributions

AEM conceived of the evaluation and drafted the quantitative portion of the methods. CRJ, TCB, and JSME drafted the introduction. CRJ developed the research agenda. HMRA, CHH, and JFME drafted the qualitative portion of the methods. All authors contributed substantially to the first draft, and read and approved the final draft.

Acknowledgements

We would like to thank the leadership of the Cluster for Reviews and Health Technology Assessments at the Norwegian Institute of Public Health for their funding of the ML team since December 2020. Thank you in particular to Department Director Rigmor C Berg, whose support was instrumental to the creation of the team and continues to enable the success of the team. Thank you also to Stijn Van De Velde and Jan Himmels, previous members of this team, for their technical expertise in building up the infrastructure for an innovation-based team.

Authors' information

AEM is a senior researcher and the machine learning team lead in the Division for Health Services at the Norwegian Institute of Public Health. She provides research and technical consultation regarding machine learning in evidence synthesis in the fields of COVID-19 and mental health to the WHO European Region. She is interested in the intersection of decolonization with evidence-based medicine and policy for marginalized populations, and the role of machine learning in this intersection.

HMRA is a researcher and the machine learning team co-lead in the Division for Health Services at the Norwegian Institute of Public Health. She specializes in qualitative evidence synthesis methodology and is very interested in how machine learning and qualitative methods intersect and the potential for cooperation. She is also an editor with the Cochrane Consumers and Communication Group.

TCB is a researcher and member of the machine learning team at Division for Health Services at the Norwegian Institute of Public Health. She has a background in nutritional epidemiology and has methodological expertise in a variety of primary and secondary research designs. She is particularly interested in new methods for the optimization of systematic review processes and communication of results, from artificial intelligence to innovative data visualization.

CHH is a researcher and a member of the machine learning in the Division for Health Services at the Norwegian Institute of Public Health. She specializes in both quantitative and qualitative evidence synthesis within health sciences, child welfare policy, and geriatric services, among others, and is interested in how machine learning can save time and optimize the review process.

JFME is a researcher and a member of the machine learning in the Division for Health Services at the Norwegian Institute of Public Health. He specializes in evidence synthesis for health decision making, and is a member of the GRADE Working Group for the study of the Evidence-to-Decision frameworks. He has worked on user experience research for decision making tools, and recently led health technology assessments as part of the European Network for Heath Technology Assessments' response to COVID-19.

CJR is a statistician at the Norwegian Institute of Public Health. He is a statistical editor for Cochrane's Effective Practice and Organisation of Care (EPOC) Review Group, supports the institute's health technology assessment work for the Norwegian specialist health services, works on randomized controlled trials, and provides statistical training and consultancy within the institute, nationally, and internationally.

5. References

Agai, E. (2020). A new machine-learning powered tool to aid citation screening for evidence synthesis: PICOPortal. *Cochrane Database of Systematic Reviews, 9*(1). doi:https://doi.org/10.1002/14651858.CD202001

Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., & Moustafa, A. A. (2019). The Application of Unsupervised Clustering Methods to Alzheimer's Disease. *Frontiers in Computational Neuroscience, 13*. doi:10.3389/fncom.2019.00031

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *WIREs Data Mining and Knowledge Discovery, 11*(5). doi:10.1002/widm.1424

Arno, A., Elliott, J., Wallace, B., Turner, T., & Thomas, J. (2021). The views of health guideline developers on the use of automation in health evidence synthesis. *Syst Rev, 10*(1), 16. doi:10.1186/s13643-020-01569-2

Bastian, H., Glasziou, P., & Chalmers, I. (2010). Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med, 7*(9), e1000326. doi:10.1371/journal.pmed.1000326

Beller, E., Clark, J., Tsafnat, G., Adams, C., Diehl, H., Lund, H., . . . founding members of the, I. g. (2018). Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Syst Rev, 7*(1), 77. doi:10.1186/s13643-018-0740-7

Berrang-Ford, L., Sietsma, A. J., Callaghan, M., Minx, J. C., Scheelbeek, P. F. D., Haddaway, N. R., . . . Dangour, A. D. (2021). Systematic mapping of global research on climate and health: a machine learning review. *The Lancet Planetary Health, 5*(8), e514-e525. doi:10.1016/s2542-5196(21)00179-0

Blaizot, A., Veettil, S. K., Saidoung, P., Moreno-Garcia, C. F., Wiratunga, N., Aceves-Martins, M., . . . Chaiyakunapruk, N. (2022). Using artificial intelligence methods for systematic review in health sciences: A systematic review. *Res Synth Methods*. doi:10.1002/jrsm.1553

Booth, A., & Carroll, C. (2015). How to build up the actionable knowledge base: the role of 'best fit' framework synthesis for studies of improvement in healthcare. *BMJ Quality & Safety, 24*(11), 700-708. doi:10.1136/bmjqs-2014-003642

Carroll, C., Booth, A., Leaviss, J., & Rick, J. (2013). “Best fit” framework synthesis: refining the method. *BMC Medical Research Methodology, 13*(1), 37. doi:10.1186/1471-2288-13-37

Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., . . . Thomas, J. (2017). Living systematic review: 1. Introduction-the why, what, when, and how. *J Clin Epidemiol, 91*, 23-30. doi:10.1016/j.jclinepi.2017.08.010

Haddaway, N. R., Callaghan, M. W., Collins, A. M., Lamb, W. F., Minx, J. C., Thomas, J., & John, D. (2020). On the use of computer-assistance to facilitate systematic mapping. *Campbell Systematic Reviews, 16*(4), e1129. doi:https://doi.org/10.1002/cl2.1129

Haddaway, N. R., Feierman, A., Grainger, M. J., Gray, C. T., Tanriver-Ayder, E., Dhaubanjar, S., & Westgate, M. J. (2019). EviAtlas: a tool for visualising evidence synthesis databases. *Environmental Evidence, 8*(1), 22. doi:10.1186/s13750-019-0167-1

Haddaway, N. R., & Westgate, M. J. (2020). Creating and curating a community of practice: introducing the evidence synthesis Hackathon and a special series in evidence synthesis technology. *Environmental Evidence, 9*(1), 28. doi:10.1186/s13750-020-00212-w

Hamel, C., Hersi, M., Kelly, S. E., Tricco, A. C., Straus, S., Wells, G., . . . Hutton, B. (2021). Guidance for using artificial intelligence for title and abstract screening while conducting knowledge syntheses. *BMC Med Res Methodol, 21*(1), 285. doi:10.1186/s12874-021-01451-2

Harrison, H., Griffin, S. J., Kuhn, I., & Usher-Smith, J. A. (2020). Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. *BMC Med Res Methodol, 20*(1), 7. doi:10.1186/s12874-020-0897-3

Hartmann, J., Wuijts, S., van der Hoek, J. P., & de Roda Husman, A. M. (2019). Use of literature mining for early identification of emerging contaminants in freshwater resources. *Environmental Evidence, 8*(1). doi:10.1186/s13750-019-0177-z

Hestevik, C., Muller, A., & Forsetlund, S. (2021). *Treatment for perpetrators of sexual violence in close relationships: a systematic review*. Retrieved from Oslo:

Himmels, J., Borge, T., Brurberg, K., & Gravningen, K. (2021). *COVID-19 and risk factors for hospital admission, severe disease and death – a rapid review, 4th update*. Retrieved from Oslo:

Himmels, J., Gomez Castaneda, M., Brurberg, K., & Gravningen, K. (2021). *COVID-19: Long-Term Symptoms after COVID-19*. Retrieved from Oslo:

Khalil, H., Tamara, L., Rada, G., & Akl, E. A. (2021). Challenges of evidence synthesis during the 2020 COVID pandemic: a scoping review. *J Clin Epidemiol, 142*, 10-18. doi:10.1016/j.jclinepi.2021.10.017

Kohl, C., McIntosh, E. J., Unger, S., Haddaway, N. R., Kecke, S., Schiemann, J., & Wilhelm, R. (2018). Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. *Environmental Evidence, 7*(1), 8. doi:10.1186/s13750-018-0115-5

Marshall, I. J., Johnson, B. T., Wang, Z., Rajasekaran, S., & Wallace, B. C. (2020). Semi-Automated evidence synthesis in health psychology: current methods and future prospects. *Health Psychol Rev, 14*(1), 145-158. doi:10.1080/17437199.2020.1716198

Marshall, I. J., Kuiper, J., Banner, E., & Wallace, B. C. (2017). Automating Biomedical Evidence Synthesis: RobotReviewer. *Proc Conf Assoc Comput Linguist Meet, 2017*, 7-12. doi:10.18653/v1/P17-4002

Marshall, I. J., Kuiper, J., & Wallace, B. C. (2016). RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inform Assoc, 23*(1), 193-201. doi:10.1093/jamia/ocv044

Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev, 8*(1), 163. doi:10.1186/s13643-019-1074-9

Miles, M. B., Huberman, A. M., & Saldana, J. (2019). *Qualitative Data Analysis: A Methods Sourcebook*: SAGE Publications.

Morichetta, A. C., Pedro Mellia, Marco. (2019). EXPLAIN-IT: Towards Explainable AI for Unsupervised Network Traffic Analysis. *Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*, 8. doi:https://doi.org/10.48550/arXiv.2003.01670

Muller, A., Ames, H., Himmels, J., Jardim, P., Nguyen, L., Rose, C., & Van de Velde, S. (2021a). *Aims and strategy for the implementation of machine learning in evidence synthesis in the Cluster for Reviews and Health Technology Assessments for 2021-2022*. Retrieved from Oslo:

Muller, A., Ames, H., Himmels, J., Jardim, P., Nguyen, L., Rose, C., & Van de Velde, S. (2021b). *Implementation of machine learning in evidence syntheses in the Cluster for Reviews and Health Technology Assessments: Final report 2020-2021*. Retrieved from Oslo:

Muller, A., Ames, H., Jardim, P., & Rose, C. (2021). Machine learning in systematic reviews: Comparing automated text clustering with Lingo3G and human researcher categorization in a rapid review. *Res Synth Methods*. doi:10.1002/jrsm.1541

O'Connor, A. M., Tsafnat, G., Gilbert, S. B., Thayer, K. A., Shemilt, I., Thomas, J., . . . Wolfe, M. S. (2019). Still moving toward automation of the systematic review process: a summary of discussions at the third meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). *Syst Rev, 8*(1), 57. doi:10.1186/s13643-019-0975-y

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst Rev, 4*(1), 5.

doi:10.1186/2046-4053-4-5

Osiński, S., Stefanowski, J., & Weiss, D. (2004). Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. In *Intelligent Information Processing and Web Mining* (pp. 359-368).

Osiński, S., & Weiss, D. (2005). *Carrot2: design of a flexible and efficient web information retrieval framework*. Paper presented at the Proceedings of the Third international conference on Advances in Web Intelligence, Lodz, Poland. https://doi.org/10.1007/11495772_68

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., . . . Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ, 372*, n71. doi:10.1136/bmj.n71

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. doi:10.1002/jrsm.1354

Rada, G., Verdugo-Paiva, F., Ávila, C., Morel-Marambio, M., Bravo-Jeria, R., Pesce, F., . . . Izcovich, A. (2020). Evidence synthesis relevant to COVID-19: a protocol for multiple systematic reviews and overviews of systematic reviews. *Medwave, 20*(3), e7868. doi:10.5867/medwave.2020.03.7867

Rosenbaum, S. (2010). *Improving the user experience of evidence : a design approach to evidence-informed health care.* (PhD). The Oslo School of Architecture and Design, Oslo, Norway. Retrieved from http://hdl.handle.net/11250/93062

Rosenbaum, S. E., Glenton, C., & Cracknell, J. (2008). User experiences of evidence-based online resources for health professionals: user testing of The Cochrane Library. *BMC Med Inform Decis Mak, 8*, 34. doi:10.1186/1472-6947-8-34

Rosenbaum, S. E., Glenton, C., Nylund, H. K., & Oxman, A. D. (2010). User testing and stakeholder feedback contributed to the development of understandable and useful Summary of Findings tables for Cochrane reviews. *J Clin Epidemiol, 63*(6), 607-619. doi:10.1016/j.jclinepi.2009.12.013

Rosenbaum, S. E., Glenton, C., Wiysonge, C. S., Abalos, E., Mignini, L., Young, T., . . . Oxman, A. D. (2011). Evidence summaries tailored to health policy-makers in low- and middle-income countries. *Bull World Health Organ, 89*(1), 54-61. doi:10.2471/blt.10.075481

Rost, T. B., Slaughter, L., Nytro, O., Muller, A. E., & Vist, G. E. (2021). Using neural networks to support high-quality evidence mapping. *BMC Bioinformatics, 22*(Suppl 11), 496. doi:10.1186/s12859-021-04396-x

Shemilt, I., Arno, A., Thomas, J., Lorenc, T., Khouja, C., Raine, G., . . . Sowden, A. (2021). Cost-effectiveness of Microsoft Academic Graph with machine learning for automated study identification in a living map of coronavirus disease 2019 (COVID-19) research [version 1; peer review: awaiting peer review]. *Wellcome Open Research, 6*(210). doi:10.12688/wellcomeopenres.17141.1

Shemilt, I., Noel-Storr, A., Thomas, J., Featherstone, R., & Mavergames, C. (2021). Machine Learning Reduced Workload for the Cochrane COVID-19 Study Register: Development and Evaluation of the Cochrane COVID-19 Study Classifier [pre-print]. *Research Square*. doi:10.21203/rs.3.rs-689189/v1

Silverman, D. (2013). *Doing Qualitative Research: A Practical Handbook Fourth Edition* (4 ed.). London: SAGE.

Stansfield, C., Thomas, J., & Kavanagh, J. (2013). 'Clustering' documents automatically to support scoping reviews of research: a case study. *Res Synth Methods, 4*(3), 230-241. doi:10.1002/jrsm.1082

Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O'Driscoll, P., & Bond, M. (2020). EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis. . London: UCL Social Research Institute.

Thomas, J., McDonald, S., Noel-Storr, A., Shemilt, I., Elliott, J., Mavergames, C., & Marshall, I. J. (2021). Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. *Journal of Clinical Epidemiology, 133*, 140-151. doi:https://doi.org/10.1016/j.jclinepi.2020.11.003

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., . . . Living Systematic Review, N. (2017). Living systematic reviews: 2. Combining human and machine effort. *J Clin Epidemiol, 91*, 31-37. doi:10.1016/j.jclinepi.2017.08.011

Thomas, J., & Stansfield, C. (2018). *Automation technologies for undertaking HTAs and systematic reviews.* Paper presented at the European Association for Health Information and Libraries (EAHIL) Conference, Cardiff, Wales.

Weisser, T., Sassmannshausen, T., Ohrndorf, D., Burggraf, P., & Wagner, J. (2020). A clustering approach for topic filtering within systematic literature reviews. *MethodsX, 7*, 100831. doi:10.1016/j.mex.2020.100831

Westgate, M. J., Haddaway, N. R., Cheng, S. H., McIntosh, E. J., Marshall, C., & Lindenmayer, D. B. (2018). Software support for environmental evidence synthesis. *Nature Ecology & Evolution, 2*(4), 588-590. doi:10.1038/s41559-018-0502-x

Figures

Figure 1

Reviews registered on PROSPERO before and after the COVID-19 pandemic began, with various ML-related terms

Legend: The amount of reviews registered on PROSPERO with various machine learning-related terms has increased by at least a factor of four from the year before COVID-19, to the second year of the pandemic. 2017 amounts were the reviews added to PROSPERO 30 Nov 2016 – 30 Nov 2017; 2019 amounts were the reviews added 30 Nov 2018 – 30 Nov 2019; and 2021 amounts corresponded to 30 Nov 2020 – 30 Nov 2021.

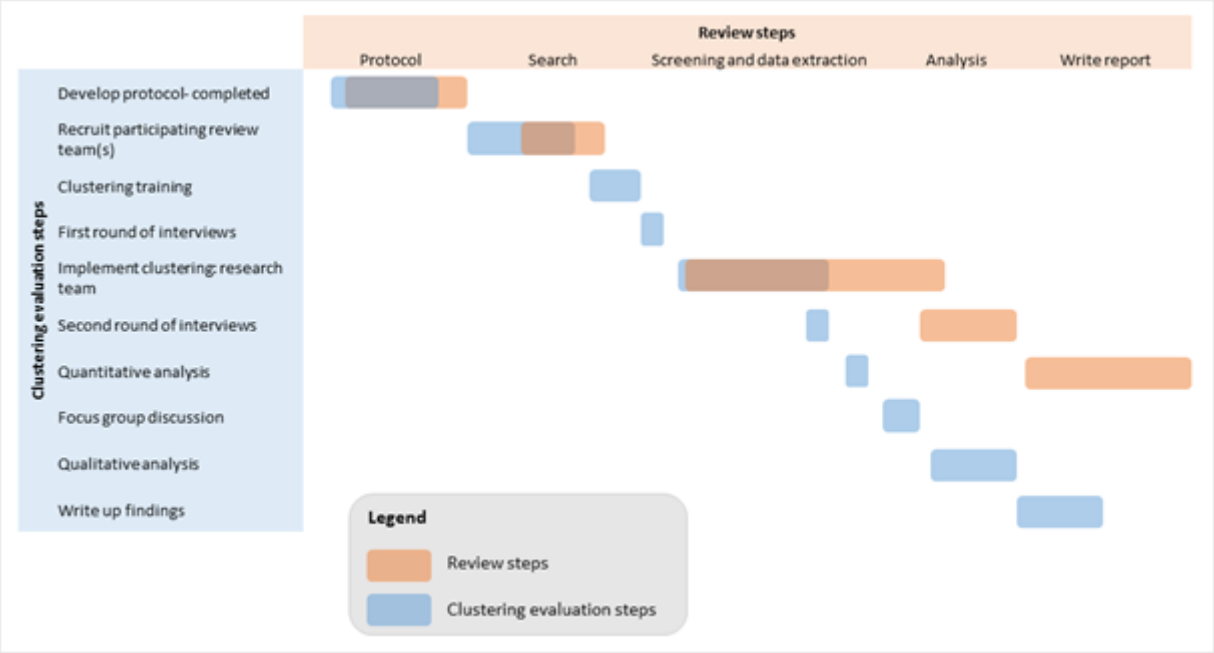

Figure 2

The steps of this evaluation mapped to the phases of the model review(s)

legend: This evaluation can be followed within the timeline of the one or two model reviews that are participating.

		Human categories (true class)		
		Population: children <18 years	Population: Young adult 18-25 years	Population: Adult >26 years
Automatically generated clusters (predicted class)	School-aged children	7	8	9
	Teenagers	1	2	3
	Youth	3	2	1

Figure 3

Sample confusion matrix

legend: A sample confusion matrix showing three predicted classes and three true classes.

		Human categories (true class)		
		Population: children <18 years	Population: Young adult 18-25 years	Population: Adult >26 years
Automatically generated clusters (predicted class)	School-aged children	0.29	0.64	0.40
	Teenagers	0.33	0.17	0.22
	Youth	0.17	0.08	0.11

Figure 4

Precision values for the above confusion matrix

legend: Precision values for the above sample confusion matrix with three predicted classes and three true classes.

- Is it better to pre-specify clustering parameters in a protocol or to plan for changing them iteratively during a review? Pre-specification in a review protocol might protect against human bias, while changing them might be necessary to obtain useable clusters.
- Does screening or data extraction aided by clustering lead to conclusions within a review or guideline that are different from those had clustering not been used?
- How can we best educate reviewers (potential users) about the mechanism behind clustering, even when the tool is user-friendly?
- How can we communicate potentials, pitfalls, and prerequisites of clustering to reviewers and to users of reviewers?

Figure 5

Practical research agenda for unsupervised clustering

legend: We suggest three directions for future research and practice around unsupervised clustering in evidence synthesis to explore. See also Muller et al. (Muller, Ames, Jardim, et al., 2021).