


Machine learning in systematic reviews: Comparing automated text clustering with Lingo3G and human researcher categorization in a rapid review

Ashley Elizabeth Muller¹  | Heather Melanie R. Ames^{1,2}  |
Patricia Sofia Jacobsen Jardim¹  | Christopher James Rose¹

¹Norwegian Institute of Public Health, Skøyen, Norway

²Cochrane Consumer and Communication Group, Centre for Health Communication and Participation, School of Psychology and Public Health, La Trobe University, Bundoora, Victoria, Australia

Correspondence

Ashley Elizabeth Muller, Norwegian Institute of Public Health, PO Box 222 0123 Skøyen, Norway.
Email: aemu@fhi.no

Abstract

Systematic reviews are resource-intensive. The machine learning tools being developed mostly focus on the study identification process, but tools to assist in analysis and categorization are also needed. One possibility is to use unsupervised automatic text clustering, in which each study is automatically assigned to one or more meaningful clusters. Our main aim was to assess the usefulness of an automated clustering method, Lingo3G, in categorizing studies in a simplified rapid review, then compare performance (precision and recall) of this method compared to manual categorization. We randomly assigned all 128 studies in a review to be coded by a human researcher blinded to cluster assignment (mimicking two independent researchers) or by a human researcher non-blinded to cluster assignment (mimicking one researcher checking another's work). We compared time use, precision and recall of manual categorization versus automated clustering. Automated clustering and manual categorization organized studies by population and intervention/context. Automated clustering failed to identify two manually identified categories but identified one additional category not identified by the human researcher. We estimate that automated clustering has similar precision to both blinded and non-blinded researchers (e.g., 88% vs. 89%), but higher recall (e.g., 89% vs. 84%). Manual categorization required 49% more time than automated clustering. Using a specific clustering algorithm, automated clustering can be helpful with categorization of and identifying patterns across studies in simpler systematic reviews. We found that the clustering was sensitive enough to group studies according to linguistic differences that often corresponded to the manual categories.

KEYWORDS

clustering, Lingo3G, machine learning, scoping reviews, systematic review

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Research Synthesis Methods* published by John Wiley & Sons Ltd.

Highlights

What is already known

Systematic reviews can use machine learning to drastically reduce time spent during study identification, particularly screening. There remains significant potential for automated approaches to reduce time needed for study analysis and categorization, but performance must also be measured.

What is new

Automated text clustering applied to included studies' titles and abstracts resulted in several useable thematic categories. The clustering algorithm Lingo3G was equally as precise as researcher categorizations, and had higher recall. Systematic reviewers without machine learning expertise can successfully implement automated text clustering.

Potential impact for RSM readers outside the authors' field

Automated text clustering can provide useable and valid categorizations of text. The time saved compared to human categorization outweighs the time needed to sort through and make sense of the automated categories.

1 | INTRODUCTION

Systematic review production is highly labor-intensive. A large number of studies must be identified and screened, and depending on the type of review, studies judged eligible must be read in full text, and their results extracted, synthesized, and reported.^{1–3} As the number of published primary studies continues to increase each year,⁴ current systematic review processes are scaling poorly: reviews are becoming more expensive to produce and more likely to require updates sooner as new studies are published. A decade ago, Bastian and colleagues⁵ reported that 11 systematic reviews were published per day and called for innovative evidence synthesis methods—although the suggestions they gave mainly involved reducing the number of primary studies conducted and published. It has also been estimated that the average intervention review takes 1.25 years to complete,² and that within 2 years of publication, one of every four systematic reviews of effect within medicine and health will become outdated.⁶ We need methods and tools that reduce unnecessary human labor and duplication of tasks to produce reviews at a speed that matches the needs of policymakers and new evidence production.

Computer-based automation and machine learning (ML) are of current interest for reducing costs and accelerating systematic review production. When successful, ML can reduce tasks that are resource-intensive (e.g., difficult or time-consuming) to tasks that can be performed more efficiently, quickly, and consistently via full- or semi-automation. Screening,^{7–9} risk of bias

assessment,¹⁰ and study design or quality classifiers^{11,12} are some of the recent applications of ML to systematic reviewing. However, systematic reviewers are often cautious when adopting new review methods and are aware that the benefits and potential harms of new methods and tools should be characterized and tested before they are adopted.¹³

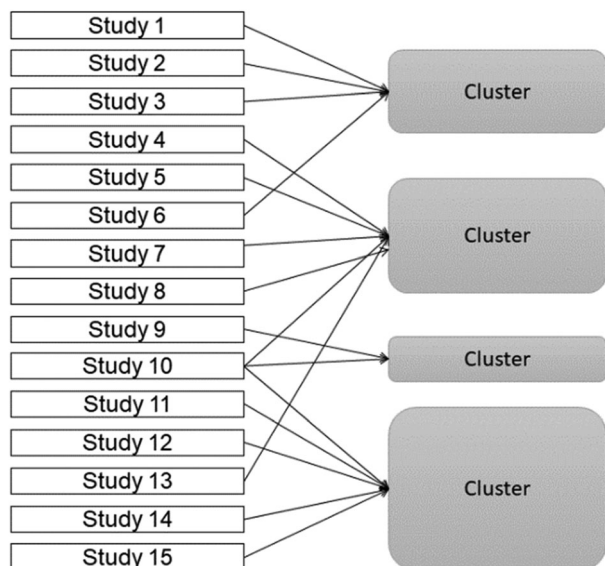
This paper addresses the problem of categorizing studies based on study content, which has application in scoping reviews that aim to map the literature published on a particular topic, population, or context, and, in some cases, identify research questions for study in subsequent systematic reviews. There is relatively little research that has applied machine learning to this problem; Stansfield and colleagues provided an early and important case study.¹⁴ This article presents a case study of the use of a clustering algorithm to define a new categorization system for a simple commissioned systematic review. We assess the utility of the resulting clusters, and compare precision, recall, and time use for completely manual categorizations versus researchers using automated clustering.

Our main aim was to assess the usefulness of an automated clustering method in categorizing studies in a simplified rapid review, then to compare performance (precision and recall) of this method against manual categorization. Ultimately, we wanted to determine whether we could “trust” a specific algorithm to cluster studies using its own categorization system, as much as we trust researchers to code to a researcher-created categorization system. Our intended audience is systematic reviewers who are not machine learning specialists.

2 | BACKGROUND

Machine learning (ML) encompasses a wide range of methods that fall under the narrower terms “supervised” and “unsupervised” learning.¹⁵ Supervised learning is a common ML approach, in which a model is first “trained” (fitted to data for which ground truth annotations are available), for the purpose of being used in a fully- or semi-supervised predictive mode to annotate new data (for which ground truth annotations are not available). In other words, a machine first “learns” how to do a new task and is then used to do that task at some performance level. Unsupervised learning can be used when ground truth annotations are not available, as in the problem we address in this paper. In this approach, a ML algorithm “learns” patterns from unannotated data to build a useful model of the population of studies from which the data originated.

Perhaps the best-known unsupervised ML method is clustering,¹⁶ in which each item in a data set is assigned to one or more automatically identified clusters such that any two data items within the same cluster are similar in some useful way, and any two clusters are dissimilar in some useful way (see Figure 1). Clustering has been used extensively in information retrieval, for example to group web search results into meaningful categories (e.g. *Aurora borealis* results separated from Aurora the singer). In some applications, clustering is hierarchical, which means that some clusters are contained within other clusters that represent higher-level concepts.



A single study can be assigned to multiple clusters, based on text in the title and abstract

FIGURE 1 Automatic clustering of text in a single hierarchy

Some clustering algorithms can also characterize each cluster in a way that is useful with respect to the domain of interest. For example, clusters can be automatically named, so that humans can understand how items within a given cluster are likely to be similar to one another. In the context of systematic reviewing, an *automated text clustering* system “analyses the distribution of terms (words) in a body of text (e.g., titles and abstracts) and identifies groups of documents that use similar combinations of words; clustering ‘engines’ often then apply a descriptive term to each cluster to aid human interpretation” (Carpentino et al. 2009, in Stansfield et al.¹⁴).

The utility of each cluster label may vary according to the algorithm’s approach: description-centric algorithms attempt to uncover descriptive, interpretable, and unambiguous names for each cluster, and then assigns text to a cluster.¹⁷ Data-centric algorithms are focused more on grouping text than providing readable cluster labels; *k*-means methods are common, which vectorize text in a bag-of-words model such that the text loses any inherent meaning. There are also algorithms that fall in between data-centric and description-centric, such as suffix text clustering, which produces cluster labels that are more adequately informative than data-centric algorithms.¹⁸

Within systematic reviews, most ML developments have addressed problems related to study identification, particularly screening. According to Marshall and Wallace,¹⁹ “machine learning systems for abstract screening have reached maturity” (p. 5). Some researchers have gone so far as to recommend ML-based screening as best practice.²⁰ Automatic data extraction and analysis represent subsequent areas of development.^{19,21} Weißer et al. have recently proposed using clustering to automatically categorize articles as low versus high interest when researchers are scoping the literature in order to develop specific research questions.²²

Clustering algorithms could also be used in the analysis phase of reviews. A simple form of analysis is categorizing studies based on content. This is important in scoping reviews, for example, in which reviewers aim to map the volume of literature that has been published on a particular topic, population, or context, and identify research questions or categories that might be studied in detail in subsequent systematic reviews. In scoping reviews, categories are informed by the research question, commissioner’s needs, data accessibility (i.e., title and abstract or full-text), and resources. Categories can be defined a priori but are often adjusted iteratively, particularly during the pilot or early phase of the process.

Systematic reviewers are unlikely to have existing categorization schemes that can be used in new reviews, or annotated sets of primary studies that would facilitate use of supervised machine learning. Such reviewers must

define a categorization scheme for each new review—either manually or, as we discuss in this article, by using ML methods such as clustering.

We are aware of only one pilot study that has used automated clustering in the systematic review process. Stansfield et al.¹⁴ retrospectively applied a description-centric clustering algorithm to two large scoping reviews and assessed the face validity of the algorithm's clusters, and the performance of the algorithm's clusters compared to researchers' manually created categories. They found that automated clusters addressed eight of nine predetermined research questions. The clustering procedure was estimated to have high precision (i.e., most of the studies assigned to a given cluster were correctly assigned to that cluster). However, relatively few studies that were actually relevant to a given cluster were assigned to it. Moreover, performance varied greatly according to the cluster. In one review, clusters adequately captured broad topics of the included studies as well as the most common interventions, but not smaller interventions. The algorithm also struggled to describe qualitative studies. Stansfield et al. examined the performance of each cluster separately but did not report summary statistics of clustering at the level of an entire review.

3 | METHODS

This experiment was an early exploration of ML within the Cluster for Reviews and Health Technology Assessments at the Norwegian Institute of Public Health. ML activities in the cluster are coordinated by the ML implementation team, of which all authors are members. A published report²³ and strategy²⁴ provide more information on completed, ongoing, and planned activities and evaluations.

3.1 | Data

This exploratory study is based on a review of the use of secure institutions for children and youth, commissioned by the Norwegian Directorate of Children, Youth and Family.²⁵ The specific product commissioned was a “systematic literature search with categorization”, a simple review product that includes only analysis of titles and abstracts.²⁶ It begins with a systematic literature search and screening of identified studies for relevance. The resulting product is an overview of the literature according to pre-defined topics (operationalized as categories), often with a focus on knowledge gaps, rather than an answer to a research question about effect or experience. It was therefore an ideal opportunity to trial

the automated text clustering function as a potential aid in sorting, categorizing, or keywording studies.

This product aimed to identify the most recent research (published 2015–2020) on the effect of secure institutions for children and youth with behavioral problems. The intervention or phenomenon of interest included secure institutions, a specific program or approach within a secure institution, or children's experiences of the effect of secure institutions. Study designs of interest were literature reviews, studies with control or comparison groups, and qualitative studies. A systematic literature search in six databases and gray literatures searches in Swedish, Norwegian, and Danish resulted in more than 13,000 references. The research team screened all studies at title/abstract level using EPPI Reviewer's “priority screening” function, a ranking algorithm that prioritizes likely relevant studies to be screened first and likely irrelevant studies to be screened last.²⁷ The product ultimately included six literature reviews, 25 controlled studies, 95 qualitative studies, and two mixed-methods studies, for a total of 128 publications. The categorization system of these 128 publications is described under *Participants and Procedures*.

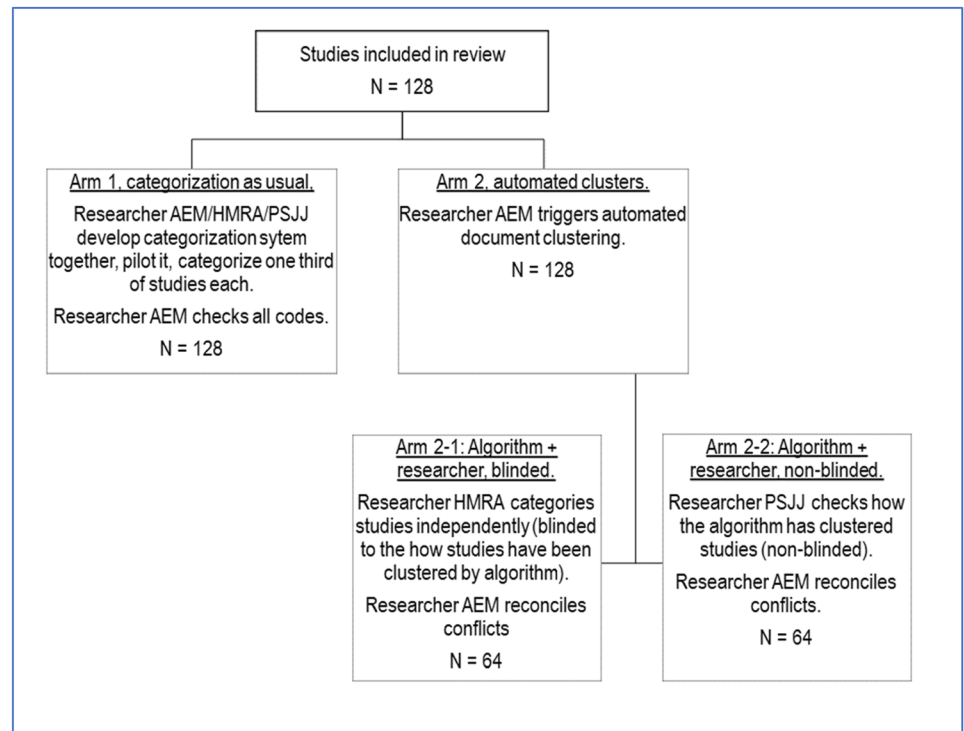
3.2 | Participants and procedures

The participants comprised two researchers with PhDs and 3–9 years' experience with systematic reviews (AEM, HMRA), and one researcher with 1 years' experience with systematic reviews (PSJJ).

We compared study categorization as per our usual practice (two human researchers, and a third to reconcile conflicts) with two ML-based approaches (Figure 2).

Arm 1 represented usual practice, and provided baseline values for time use and precision/recall. Three researchers (AEM, HMRA, and PSJJ) created a coding system, and each researcher applied the system to a distinct subset of all included studies; one of the researchers (AEM) then checked and reconciled all categorizations. The categorization system used in arm 1 was a two-level system created by the lead researcher (AEM) and refined through discussion with the other two researchers (HMRA and PSJJ). The categories were defined in terms of study design, context/intervention (a variable that can be applied to describe both the intervention tested in experiments or the setting and topic explored in qualitative studies), population, and country. These variables are typically delivered to this commissioner as an output of this type of review. Sub-categories were created through discussion among the three researchers and were piloted for usefulness and to reduce ambiguity. The final coding for arm 1 was agreement of two researchers.

FIGURE 2 Assignment of the 128 studies in human-only and algorithm-assisted arms [Colour figure can be viewed at wileyonlinelibrary.com]



Arm 2 used ML-based approaches. Automated clustering was triggered by one of the researchers (AEM), which automatically created a coding system and applied it to categorize all 128 included studies. The categorization system used in arms 2-1 and 2-2 was defined by applying EPPI Reviewer's²⁷ built-in text clustering function, which uses the Lingo3G clustering engine powered by Carrot Search.^{28,29} As a description-centric algorithm, the Lingo3G website consistently highlights the instantaneous utility and meaningfulness of cluster labels in its product description: “clearly-labeled” folders enable “instant analysis” and give the user an “instant overview” of text, and will help the user focus on “specific subject(s)”.³⁰ Lingo3G's focus on informative labels and user understanding differs from more common data-centric clustering algorithms.

The default clustering settings displayed in EPPI Reviewer are two hierarchy depths, a minimum cluster size of 10%, and a maximum cluster size of 35%. We changed these parameters iteratively to obtain immediately sensible clusters and proceeded with non-hierarchical clustering. We retained the minimum cluster size of 10% and increased the maximum cluster size to 50%. While the automated clustering procedure provides cluster names, we found it necessary to edit some of the names suggested by the software to be more easily understood by other researchers. These names were chosen by reviewer AEM, by studying the titles and abstracts of the studies assigned to the poorly named clusters. Clusters

that were not judged as useful after exploration were discarded.

We randomized studies in a 1:1 ratio across arms 2-1 and 2-2 using EPPI Reviewer's random distribution function, resulting in 64 different studies in each arm.

Arm 2-1 assessed the validity of the automatic clusters by having a researcher (HMRA) apply the automatically generated coding system to studies, blinded to how the algorithm had clustered them. Another researcher (AEM) then checked and reconciled all categorizations (as per usual practice).

Arm 2-2 assessed how a researcher who was not blinded to the clustering algorithm would categorize studies. A second researcher (PSJJ) simply checked how the clustering algorithm categorized the included studies, as she would check another researcher's data extraction.

Note that the ML-based approaches only require two researchers rather than three, as per usual practice, as the algorithm itself represented a third researcher. However, for the purpose of comparing the two approaches, it was necessary that the human tasks were performed by different people, that is, researchers HMRA and PSJJ.

Both parts of Arm 2 were completed after Arm 1, and Arm 1 was the commissioned review itself. By the completion of the review, all researchers were familiar with all studies. There was no way to avoid their knowledge, given that a manual categorization process beginning with a new categorization system requires in-depth knowledge of included studies.

3.3 | Analysis

To assess usefulness, one researcher (AEM) mapped each automatically generated cluster to a manually generated category. This provided a simple visualization of overlap and gaps between the two approaches. Our main aim was to objectively assess performance of the automated clustering method, to determine whether we could “trust” an algorithm to cluster studies using its own categorization system, as much as we trust researchers to code to a researcher-created categorization system. We therefore computed precision and recall (see Appendix 1) with respect to the final coding,³¹ treating both the algorithm and human researchers as coders/researchers. Finally, we recorded and compared the time spent coding using automated- versus human-generated categories. Each researcher recorded her time manually in an Excel file, for each step and task, and we calculated the total time used for each arm.

4 | RESULTS

Figure 3 shows the 16 clusters identified by EPPI Reviewer's document clustering function on the left side. The right side displays the conceptual re-organization of 12 of these clusters into a two-level hierarchy (chosen for ease of comparison to the manually created categories). Of the original 16 clusters, four clusters (young people, sample, suggest, and conclusion) were judged to be irrelevant upon examination and were discarded. One cluster (no relevant categories/no abstract) corresponded to the

eight studies that either did not have an abstract or were not assigned to any of the other 15 clusters. This cluster contained nine studies; seven of which were identified through gray literature searches, lacked abstracts, and were published in Norwegian or Swedish. These languages are among the 19 languages that Lingo3G can automatically detect and process, and therefore could have been included in the other clusters had they contained abstracts.

4.1 | Usefulness of automated clustering

Figure 4 displays the content of the automated clusters and manually created categories, with overlapping categories highlighted in yellow. Both approaches contained categories that described contexts/interventions and populations. Within the contexts/interventions category, both approaches identified when a study focused on a program or approach within a secure institution (e.g., anger management, animal therapy) rather than on the secure institution itself.

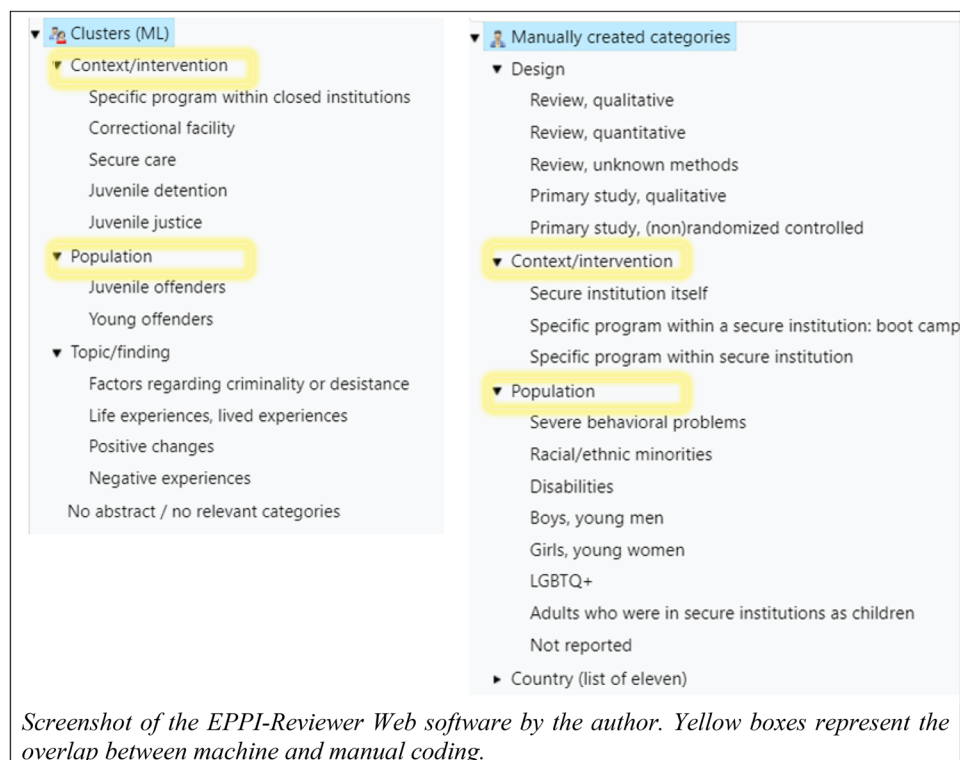
The figure shows there were no automatically generated clusters that correspond to two of the manually created top-level categories, namely country and study design, which were of interest to the commissioners. Table 1 below shows the amount of studies confirmed to be in each automated cluster or human-created category, after human coding and agreement. The lack of automated country codes might be explained by the fact that only 73 (57%) of the 128 studies specified country in the abstract and could be manually categorized. In addition,

Clusters before sorting	Clusters after sorting into two levels
<ul style="list-style-type: none"> ▼ Lingo3G clusters Juvenile justice Young people Juvenile offenders Correctional facilities Young offenders Secure care Negative experiences Positive changes Sample Factors regarding criminality/desistance Suggest Specific program within institution Life experiences, lived experiences Conclusion No relevant categories / no abstracts Juvenile detention 	<ul style="list-style-type: none"> ▼ Clusters (ML) ▼ Context/intervention <ul style="list-style-type: none"> Specific program within closed institutions Correctional facility Secure care Juvenile detention Juvenile justice ▼ Population <ul style="list-style-type: none"> Juvenile offenders Young offenders ▼ Topic/finding <ul style="list-style-type: none"> Factors regarding criminality or desistance Life experiences, lived experiences Positive changes Negative experiences No abstract / no relevant categories

Screenshot of the EPPI-Reviewer Web software by the author.

FIGURE 3 Making sense of categories created through automated clustering [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 4 Comparison of automated clusters and manually created categories [Colour figure can be viewed at wileyonlinelibrary.com]



nine of 11 countries were reported by less than 10% of studies; our 10% minimum cluster size cut-off would therefore have prevented studies being clustered around these rarer words.

One reason for a lack of automated clusters relating to study design is that it may be challenging for this unsupervised ML method to infer important differences in study design. However, the clustering algorithm identified categories related to study topics and findings (factors related to criminality/desistance; life experiences, lived experiences; positive changes; and negative experiences), which were not part of the manual categorization scheme. The topic/finding top-level category corresponded roughly to study design. Sixty-one (62%) of the 98 qualitative studies received a topic-finding category, most often “life experiences, lived experiences” and “negative experiences”. Only 12 (38%) of the 32 quantitative studies received a topic/finding category, most often “factors related to criminality/desistance”.

“Population” was a top-level category in both automated clusters and manually generated categories. While researchers did not deem it useful in their manually created categories to group studies according to semantic differences such as “juvenile” or “young”, it was straightforward to place studies in one of these two categories, as study authors used either one or the other phrase to describe their populations.

4.2 | Performance of automated clustering (using algorithmically-generated categories)

The leftmost column of Table 2 shows the cumulative performance of the researchers using manually created categories (Arm 1). Both precision and recall of the researchers exceeded 96%, likely due to how these categories were created through discussion among the three researchers, were piloted, and were intended to be unambiguous and mainly mutually exclusive.

Table 2 also shows estimates of precision and recall for researchers and automated clustering in which consensus between two researchers was used for final coding. In Arm 2-1, in which automated clustering represented one independent researcher’s coding, automated clustering and the actual researcher’s precision rates were similar: 81%–82% of their codes identified relevant studies. In this study, recall for automated clustering was 10% points greater than for the researcher. The statistical analysis suggests it is plausible that recall may be either similar across the two approaches or that the automated approach may be superior.

In Arm 2-2, in which one researcher saw and checked the clusters rather than coding blind, precision was again identical for both the algorithm and the researcher, and higher than in Arm 2-1. The clustering algorithm again had better recall than the researcher, and in Arm 2-2,

TABLE 1 Studies assigned to automated clusters and manually created categories, after researcher agreement

Automated clusters	N	Manually created categories	N
Context/intervention		Design	
Specific programs within closed institutions	42	Review, qualitative	1
Institution name: correctional facility	23	Review, quantitative	5
Institution name: secure care	23	Review, unknown methods	0
Institution name: juvenile detention	22	Primary study, qualitative	97
Institution name: juvenile justice	28	Primary study, (non-)randomized controlled	27
Population		Context/intervention	
Juvenile offenders	16	Secure institution itself	102
Young offenders	14	Specific programs within secure institutions: boot camp	2
		Specific programs within secure institutions	25
Topic/finding		Population	
Factors regarding criminality or desistance	26	Severe behavioral problems	10
Life experiences, lived experiences	40	Racial/ethnic minorities	9
Positive changes	13	Disabilities	1
Negative changes	18	Boys, young men	31
		Girls, young women	28
		LGBTQ+	2
		Adults who were in secure institutions as children	12
		Not reported	27
		Country	
		Canada	1
		Denmark	6
		Norway	2
		Israel	1
		Poland	1
		Portugal	2
		Sweden	9
		Netherlands	6
		Great Britain	13
		USA	33
		Not reported	55

TABLE 2 Precision and recall of human-created categories and automated document-clustering categories (with 95% confidence intervals)

	Arm 1: Coding using human-created categories		Arm 2: Coding using algorithm's clusters			
	Researcher + researcher (non-blinded)		Arm 2-1: Algorithm + researcher (blinded)		Arm 2-2: Algorithm + researcher (non-blinded)	
	Precision	Recall	Precision	Recall	Precision	Recall
Clustering algorithm	–	–	0.825 (0.747–0.883)	0.780 (0.699–0.843)	0.884 (0.813–0.931)	0.843 (0.768–0.896)
Researcher	0.989 (0.968–0.998)	0.965 (0.937–0.982)	0.810 (0.723–0.874)	0.669 (0.583–0.745)	0.890 (0.816–0.937)	0.770 (0.689–0.835)

TABLE 3 Time used in each arm (hours)

Steps	Arm 1: Coding using human-created categories	Arm 2: Coding using algorithm's clusters	
		Arm 2-1: Algorithm + researcher (blinded)	Arm 2-2: Algorithm + researcher (non-blinded)
a) Making categories	0.5	0.01	
b) Making sense of automated categories	–	3.45	
c) Coding and coming to agreement	10.9	2.0	2.2
Total time used	11.4	7.66	

retrieved 84% of relevant studies, compared to the researcher's recall of 77%.

Table 3 displays the time spent in each arm. In Arm 1, manually creating codes, piloting them together, applying codes, and reconciling conflicts required 11.4 h (approximately 49% longer than the automated approaches). The majority of this time was spent in applying and reconciling coding. In Arm 2, the time needed to run the clustering algorithm, interpret the clusters, apply the clusters independently (Arm 2-2) or check the algorithm's clusters (Arm 2-1), and reconcile conflicts was 7.7 h for the 128 studies. Almost half of this time—3.45 min—was spent in making sense of the clusters, including re-naming ambiguous clusters and discarding irrelevant clusters. Coding using the algorithm's clusters took less than 40% of the time that coding using human-created categories did (4.2 h compared to 10.9 h).

5 | DISCUSSION

In this exploratory validation study, we tested the usefulness of automated clustering in categorizing 128 studies in a simplified systematic review. We assessed the performance of both this description-centric algorithm, Lingo3G, and two researchers—blinded and non-blinded—against final coding decisions and compared performance and resource use of categorizing with help of the algorithm against categorizing manually. Clustering provided useful categories for the review, but these were not exhaustive; it could not have replaced researcher-created categories. In terms of performance, the algorithm had remarkably similar precision to any one experienced systematic review researcher when assessing both against final coding, and 7%–11% better recall than any one researcher. The automated approach also used 33% less time. We therefore see exciting potential to supplement researcher categorization with automated clustering, and our study provides evidence that such methods can be as accurate as one or two researchers.

There were surprisingly helpful overlaps between the automated clusters and manual categories, as well as clear benefits to each of the approaches; see Table 4. The clustering algorithm was unable to organically cluster

TABLE 4 Summary of human-created categorization and automated clustering benefits and limitations

	Human-created categorization	Automated clustering
Benefits	<p>Can create categorization systems with specific and complicated structures, such as different levels of hierarchies, requiring mutual exclusivity, and so forth</p> <p>Potentially more trusted than automated clustering by commissioners</p> <p>Can identify “empty” categories, that is, knowledge gaps or a lack of studies that fit into a category of interest</p>	<p>Negligible time needed to create the clusters</p> <p>Pilot testing is not necessary</p> <p>Can be trusted to perform as well as a researcher</p> <p>Highlights breadth/range of clusters</p> <p>May capture topics not identified by researchers</p> <p>Range of flexible settings</p>
Limitations	<p>Time-consuming to create the categorization system, pilot test, categorize, and check others' categorizations</p>	<p>Cannot be used exclusively to categorize to a pre-determined categorization system</p> <p>Some time and interpretation needed to make sense of some clusters</p>

studies according to a pre-determined categorization system. However, it was sensitive to linguistic differences that sometimes corresponded to pre-determined categories. In social and welfare evidence synthesis we are often faced with summarizing effects of interventions and policies that lack internationally agreed upon names and definitions. In this project, when the researchers manually coded interventions/contexts and populations, they intentionally disregarded what they assessed to be country-level variation in terms, in a more semantic style of categorization due to variation in intervention and policy naming. For example, a study reporting on youth in “secure care” in the UK and a study reporting on youth in “juvenile detention facilities” in the United States were manually coded to Context/intervention > Secure institution itself. The exact name was not important in the manually created categories. However, the clustering algorithm honed in on these linguistic differences, and these two studies were clustered to Context/intervention > Secure care, and Context/intervention > Juvenile detention, respectively, which also corresponded to the manual country codes of the UK and USA. In a subsequent addition to the project, the commissioner requested studies divided by type of secure institution. The automated document clustering categories provide exactly those groups, saving us the time it would have taken to recode from scratch.

One clear benefit of the algorithm was that it created a unique cluster—Topics/Findings—that did not have a manually created counterpart, and that proved particularly useful. After the original simplified review was delivered, the commissioner requested extensive summaries of the six identified literature reviews, with a focus on their topics and themes, and particularly whether results indicated positive or negative effects of secure institutions. Researchers were able to refer to this cluster's sub-categories as they summarized these publications.³² The algorithm therefore proved useful as a supplement, but not a replacement.

Automated clustering required significantly less time, even accounting for the three and a half hours needed to make sense of clusters, which included re-labelling some and discarding others. In fact, making sense of the clusters was the most time-intensive step. Clusters' names tended to be the words that characterized the cluster. This was different than usual practice of categorizing according to study content for two reasons: first, researchers often attempt to create mutually exclusive categories or at least minimally overlapping categories, while document clustering does not allow for this. Second, researchers often create categories within the same conceptual “plane” as one another: mutually exclusive types of program designs, mutually exclusive population groups, and mutually exclusive contexts, rather than a category that describes a particular population *and* a particular context *and* a particular program design.

Automated clustering is just as likely to cluster studies once into population groups and again into contexts, meaning there will not only be overlapping categories, with the same study categorized into a program design-related cluster and into a population-related cluster, but the categories may represent different “planes”.

Overall, this suggests that automated clustering has limited utility, and does not save time, in assigning studies into *pre*-determined categories that may be hierarchically organized or with rules such as mutual exclusivity. Rather, the unsupervised nature of clustering points to its usefulness in highlighting similarities between studies. Stansfield et al.¹⁴ also reported that Lingo3G succeeded in accurately describing a wider range of content than human categorization but could not cluster according to all pre-defined categories. A major advantage of description-centric algorithms such as Lingo3G over standard clustering algorithms that use a bag-of-words approach is their production of reasonable, immediately understandable cluster labels—nevertheless, we spent more than 3 h making sense of them. We therefore assume that this stage would have required even more time had we used a data-centric algorithm. Time savings may have been greater had we used a different algorithm that we were able to fine-tune more, then apply to new data.

In addition to requiring less time, the clustering algorithm performed as well as any two researchers categorizing according to the algorithm's system, whether blinded or non-blinded. While there have been studies exploring clustering algorithms within evidence synthesis for comparison, our findings of the algorithm's precision (83%–88%) were similar to the precision reported by the algorithm's developers in their initial user study (80%–95%).²⁸ It is possible that the range of precision and recall could have been related to differences in the two arms' studies, although we hope that randomizing studies protects against systematic differences. We also saw no indication of confirmation bias in the arm in which a researcher was not blinded to the algorithm's assignment of studies. We interpret these results to mean that this particular clustering algorithm's “decisions” regarding how studies relate to each other can be trusted as much as when researchers themselves decide how studies relate to each other.

Performance of this clustering algorithm is based first and foremost on researcher acceptance of automated categories, and second on recall/precision of the accepted automated categories, compared to researcher classification. The act of accepting (or rejecting, or modifying) algorithmically generated clusters represents human input and intervention into ML tools. We suggest this human engagement be regarded as a necessary step in implementing ML tools in evidence synthesis, even when those tools could allow for full automation.

5.1 | Recommendations for evidence synthesis

In a systematic scoping review or a systematic literature search with categorization, automated clustering using the description-centric algorithm Lingo3G should be used to create initial categories, before manually creating a categorization system. All researchers involved in this process should carefully review clusters for relevance and clarity. Only the clusters that are useful should be carried forward; ambiguous clusters or those that are taking time to understand should be discarded. As automated clustering may outperform human researchers with respect to recall (as our study suggests), we can probably depend on it to identify more studies than a researcher who codes blind. One researcher may then check the studies coded to these selected categories for accuracy. Although we see no evidence in this study that a researcher will be more precise than document clustering, we hypothesize that researcher precision will increase if all researchers are involved in assessing and understanding the automated clusters. It may be useful for a researcher to check the precision of document clustering categories. These hypotheses should be explored in subsequent studies.

After reviewing the automated clusters, researchers should manually create and code any supplemental categories as needed. This is similar to a best-fit framework synthesis^{33,34} used in qualitative evidence syntheses, in which authors categorize data into a pre-existing framework. Any data that are left outside of the framework are then thematically analyzed to create new framework areas. The framework is then expanded to accommodate the new areas creating a new framework that includes all of the relevant data. By using a hybrid human- and automated-categorization system, future reviews may benefit from the resources saved by automation as well as the specificity provided by manual categorization.

The automated clusters were extremely helpful in a subsequent, smaller commission. We echo Stansfield et al.'s¹⁴ recommendation that automated clustering could help provide direction and focus in a large review, when there is a need to create a smaller dataset. Applications for automated clustering may therefore exist in larger systematic reviews with specific quantitative or qualitative research questions, or in preliminary searches for such reviews.

5.2 | Study strengths and limitations

Our findings are based on a single review, only one clustering algorithm, and three researchers. More comprehensive prospective studies and utilizing different

clustering algorithms would be required to provide rigorous comparisons of human and automated approaches. There are certainly more sophisticated algorithms to explore. Another more advanced approach could use language models that perform on more conceptual than tokenist levels, such as the Generative Pre-trained Transformer 3 model with its 175 billion language parameters.³⁵ At the same time, this particular algorithm was user-friendly and available in a popular systematic review software. The most cutting-edge and complex ML systems are often the least user-friendly and transparent, and both characteristics undermine uptake of ML in evidence synthesis¹³ as well as more broadly.³⁶ Systematic reviewers are more likely to accept a ML tool if it is interpretable, as Lingo3G's clusters were. We expect more sophisticated algorithms to become more user-friendly for systematic reviewers in the future.

The research reported in this paper was carried out during a commissioned review that had a short time frame, a large number of search hits, and a large number of relevant studies. We are unsure of how well automated clustering would work on a review with a limited number of included studies. The timesavings in that scenario would be limited and potentially not worthwhile. Our time estimates are also likely dependent upon the clustering algorithm we used; different algorithms may require more or less time to interpret labels and discard irrelevant clusters.

5.3 | Future research agenda

We believe this is the first study in this area and hope our work is a useful contribution that can be used to help plan more rigorous randomized studies. Adoption of ML methods are gaining traction within evidence synthesis—for example, the well-known PRISMA study flow templates for systematic reviews now include specification of manual versus automated study identification,³⁷ and recent reviews have further tailored PRISMA figures to include neural network-based knowledge graphs^{38,39}—but these are still the exceptions, rather than the rule. Research needs to explore how ML, particularly unsupervised tools with modifiable parameters, should be handled in the protocol stage: Is it better to pre-specify parameters in a study protocol, thereby protecting against human bias in modifying parameters in a particular direction, or to plan for changing parameters iteratively in order to obtain sensible clusters? Do ML methods lead to conclusions within a systematic review, and ultimately in a guideline, different from those had ML methods not been used? Finally, how can we best educate systematic reviewers and other users about the mechanism behind ML tools, even when the tool is user-friendly, as well as

about potential consequences and trade-offs occurring when automating previously manual tasks?

One way in which the risks and benefits of clustering and other ML-based tools could be studied is with a prospective case-control study of systematic reviews and health technology assessments with the same or similar inclusion criteria. By pairing reviews and health technology assessments that did and did not use ML, it should be possible to analyze outcomes such as time-to-publication and human agreement in data extraction. Randomizing a pre-specified amount of commissioned reviews to use or not use ML tools could also provide comparative data about conclusions.

6 | CONCLUSION

This study shows that it is feasible and can be useful to use automated clustering to create, inform, or otherwise supplement study categorization systems for scoping reviews or more simplified systematic review products. We estimated that automated clustering with the description-centric Lingo3G algorithm is as precise as human researcher categorization and uses 33% less time. Coding to human-created categories took far more time than coding to clusters, but there was a sunk cost of almost 3.5 h in making sense of the clusters, even using an algorithm intended on providing descriptive and meaningful cluster labels. At the same time, the clusters identified by machine learning did not include essential categories such as country or study design. In the future, review teams could begin the categorization process by applying a clustering algorithm to the included studies. These clusters should be examined and discussed within the research team and ambiguous or unnecessary clusters removed. The remaining clusters could then be used as the foundation for further categorization of the included studies, and researchers can trust the performance of the algorithm as much as one another's. Importantly, we suggest that this particular clustering algorithm, available in a popular systematic review software, can be used by systematic reviewers who are not machine learning experts.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Ashley Elizabeth Muller was the project leader, conceived of the experiment, and collected data. Ashley Elizabeth Muller, Patricia Sofia Jacobsen Jardim, and Heather Melanie R. Ames designed and conducted the experiment. Christopher James Rose analyzed the data.

All authors contributed substantially to the final draft and have approved its submission.


DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Ashley Elizabeth Muller  <https://orcid.org/0000-0001-7819-6697>

Heather Melanie R. Ames  <https://orcid.org/0000-0001-8509-7160>

Patricia Sofia Jacobsen Jardim  <https://orcid.org/0000-0001-6457-8168>

REFERENCES

1. Nussbaumer-Streit B, Ellen M, Klerings I, et al. Resource use during systematic review production varies widely: a scoping review. *J Clin Epidemiol.* 2021;139:287-296. doi:10.1016/j.jclinepi.2021.05.019
2. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open.* 2017;7(2):e012545. doi:10.1136/bmjopen-2016-012545
3. Allen IE, Olkin I. Estimating time to conduct a meta-analysis from number of citations retrieved. *JAMA.* 1999;282(7):634-635. doi:10.1001/jama.282.7.634
4. Bornmann L, Mutz R. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *J Assoc Inf Sci Technol.* 2015;66(11):2215-2222. doi:10.1002/asi.23329
5. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med.* 2010;7(9):e1000326. doi:10.1371/journal.pmed.1000326
6. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. *Ann Intern Med.* 2007;147(4):224-233. doi:10.7326/0003-4819-147-4-200708210-00179
7. Przybyla P, Brockmeier AJ, Kontonatsios G, et al. Prioritising references for systematic reviews with RobotAnalyst: a user study. *Res Synth Methods.* 2018;9(3):470-488. doi:10.1002/jrsm.1311
8. Shemilt I, Simon A, Hollands GJ, et al. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Res Synth Methods.* 2014;5(1):31-49. doi:10.1002/jrsm.1093
9. Callaghan MW, Muller-Hansen F. Statistical stopping criteria for automated screening in systematic reviews. *Syst Rev.* 2020;9(1):273. doi:10.1186/s13643-020-01521-4
10. Armijo-Olivo S, Craig R, Campbell S. Comparing machine and human reviewers to evaluate the risk of bias in randomized controlled trials. *Res Synth Methods.* 2020;11(3):484-493. doi:10.1002/jrsm.1398
11. Langlois A, Nie JY, Thomas J, Hong QN, Pluye P. Discriminating between empirical studies and nonempirical works using automated text classification. *Res Synth Methods.* 2018;9(4):587-601. doi:10.1002/jrsm.1317

12. Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. *Res Synth Methods*. 2018;9(4):602-614. doi:10.1002/jrsm.1287
13. Arno A, Elliott J, Wallace B, Turner T, Thomas J. The views of health guideline developers on the use of automation in health evidence synthesis. *Syst Rev*. 2021;10(1):16. doi:10.1186/s13643-020-01569-2
14. Stansfield C, Thomas J, Kavanagh J. 'Clustering' documents automatically to support scoping reviews of research: a case study. *Res Synth Methods*. 2013;4(3):230-241. doi:10.1002/jrsm.1082
15. MBAJL RB. Frameworks for scaling up machine learning. In: Ron Bekkerman MB, Langford J, eds. *Scaling up Machine Learning: Parallel and Distributed Approaches*. Cambridge University Press; 2012.
16. Aggarwal CC, Zhai C. A survey of text clustering algorithms. In: Aggarwal CC, Zhai C, eds. *Mining Text Data*. Springer; 2012:77-128.
17. Kozłowski M, Rybicki H. Clustering of semantically enriched short texts. *J Intell Inf Syst*. 2018;53(1):69-92. doi:10.1007/s10844-018-0541-4
18. Carpineto C, Osinski S, Romano G, Weiss D. A survey of web clustering engines. *ACM Comput Surv* 2009;41:17:1-17:38.
19. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev*. 2019;8(1):163. doi:10.1186/s13643-019-1074-9
20. Polanin JR, Pigott TD, Espelage DL, Grotperter JK. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Res Synth Methods*. 2019;10(3):330-342. doi:10.1002/jrsm.1354
21. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. *Syst Rev*. 2015;4:78. doi:10.1186/s13643-015-0066-7
22. Weisser T, Sassmannshausen T, Ohrndorf D, Burggraf P, Wagner J. A clustering approach for topic filtering within systematic literature reviews. *MethodsX*. 2020;7:100831. doi:10.1016/j.mex.2020.100831
23. Muller AE, Ames HMR, Himmels JPW, et al. *Implementation of Machine Learning in Evidence Syntheses in the Cluster for Reviews and Health Technology Assessments: Final Report 2020-2021*. Norwegian Institute of Public Health; 2021:83.
24. Muller AE, Ames HMR, Himmels JPW, et al. *Aims and Strategy for the Implementation of Machine Learning in Evidence Synthesis in the Cluster for Reviews and Health Technology Assessments for 2021-2022*; 2021:83.
25. Muller AE, Jardim PSJ, Ames HMR, Zinöcker S. *Secure Institutions for Youth: Systematic Literature Search with Categorization*. Norwegian Institute of Public Health; 2020:42.
26. Norwegian Institute of Public Health. *Handbook of evidence synthesis [Slik oppsummerer vi forskning. Håndbok for Folke helseinstituttet]*. 2018. <https://www.fhi.no/kk/oppsummert-forskning-for-helsetjenesten/hva-er-en-kunnskapsoppsummering/>
27. EPPI-reviewer: advanced software for systematic reviews, maps and evidence synthesis. UCL Social Research Institute; 2020. <https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2967>
28. Osiński S, Stefanowski J, Weiss D. Lingo: search results clustering algorithm based on singular value decomposition. *Intell Inf Process Web Min*. 2004;359-368.
29. Osiński S, Weiss D. Carrot2: design of a flexible and efficient web information retrieval framework. Presented at: Proceedings of the Third International Conference on Advances in Web Intelligence; Lodz, Poland. 2005 doi: 10.1007/11495772_68
30. CarrotSearch. Lingo3G or Carrot2? Updated 01.01.2021. Accessed 01 June, 2021. <https://carrotsearch.com/lingo3g/comparison/>
31. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432
32. Muller A, Jardim P, Ames H. *Secure Institutions for Youth: Six Synopses*. Norwegian Institute of Public Health; 2020:20.
33. Shaw L, Nunns M, Briscoe S, Anderson R, Thompson CJ. A "rapid best-fit" model for framework synthesis: using research objectives to structure analysis within a rapid review of qualitative evidence. *Res Synth Methods*. 2020. doi:10.1002/jrsm.1462
34. Carroll C, Booth A, Leaviss J, Rick J. "Best fit" framework synthesis: refining the method. *BMC Med Res Methodol*. 2013;13(1):37. doi:10.1186/1471-2288-13-37
35. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantam A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. Arxiv preprint. arXiv:2005.14165. 2020:p. 75.
36. USACM. *Statement on Algorithmic Transparency and Accountability*. Association for Computing Machinery US Public Policy Council; 2017:2.
37. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71
38. Havnes IA, Muller AE. Nonprescribed androgen use among women and trans men. *Curr Opin Endocrinol Diabetes Obes*. 2021;28(6):595-603. doi:10.1097/med.0000000000000680
39. JPW H, TC B, KG B, KM G. *Covid-19 and Risk Factors for Hospital Admission, Severe Disease and Death—a Rapid Review, 4th Update*. 2021. 40. <https://fhi.brage.unit.no/fhi-xmlui/handle/11250/2757595>

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Muller AE, Ames HMR, Jardim PSJ, Rose CJ. Machine learning in systematic reviews: Comparing automated text clustering with Lingo3G and human researcher categorization in a rapid review. *Res Syn Meth*. 2022;13(2):229-241. doi:10.1002/jrsm.1541