Journal of the Medical Library Association

www.mlanet.org

Vol. 113 No.1 (2025) | January 2025

Journal of the Medical Library Association

Volume 113
Number 1
January 2025
ISSN: 1558-9439 (Online)
113(1):1-112(2025)
J Med Libr Assoc

The *Journal of the Medical Library Association (JMLA)* (ISSN: 1536-5050) is the official journal of the Medical Library Association (MLA) and is published four times a year, in January, April, July, and October.

Manuscripts: Instructions for authors can be found at jmla.mlanet.org/ojs/jmla/about/submissions. Manuscripts intended for peer review should be submitted at jmla.mlanet.org. Other communications should be sent to jmla@journals.pitt.edu.

Contributions are evaluated by a double-blind process, using reviewers from the *JMLA* Editorial Board and/or subject experts.

Opinions expressed here are those of the authors and do not necessarily represent those of the association.

JMLA is a continuation of the *Bulletin of the Medical Library Association (BMLA),* beginning with v. 90, 2002. The *BMLA* began with n.s. v. 1 in 1911 as the successor to the *Medical Library and Historical Journal* (Association of Medical Librarians), v. 1–5, 1903–1907; the *Journal of the Association of Medical Librarians,* v. 1, 1902; and *Medical Libraries* (Association of Medical Librarians), v. 1–5, 1898–1902 (official organ through 1901).

Articles from *JMLA* are indexed or abstracted in the following publications and databases: Bibliography of the History of Medicine, Cochrane Methodology Register, CINAHL, Current Awareness in Biological Sciences, Current Work in the History of Medicine, EMBASE/Excerpta Medica, Health and Psychosocial Instruments, Index Medicus, Index to Dental Literature, Information Science Abstracts, International Nursing Index, Könyvtári és Dokumentációs Szakirodalm, Library and Information Science Abstracts, and Library Literature & Information Science.

WEBSITE: The *JMLA* website with full content beginning in 2016 is at jmla.mlanet.org.

DIGITAL ARCHIVES: Archives of *JMLA* are available at www.pubmedcentral.nih.gov.

EDITORIAL

DOI: dx.doi.org/10.5195/jmla.2025.2099

Revisiting *JMLA* case reports: a publication category for driving innovation in health sciences librarianship

Jill T. Boruff, AHIP; Michelle Kraft, AHIP, FMLA; Alexander J. Carroll, AHIP

See end of article for authors' affiliations.

In the April 2019 issue (Vol. 106 No. 3), the *Journal of the Medical Library Association (JMLA)* debuted its Case Report publication category. In the years following this decision, the Case Reports category has grown into an integral component of *JMLA*. In this editorial, the *JMLA* Editorial Team highlights the value of case reports and outlines strategies authors can use to draft impactful manuscripts for this category.

INTRODUCTION

In the April 2019 issue (Vol. 106 No. 3), the *Journal of the Medical Library Association (JMLA)* debuted its Case Report publication category [1]. The Case Report category replaced the preceding Case Study category in efforts to delineate the difference between case reports as a publication category and case study as a research method [2]. In the years following this decision, the Case Reports category has grown into an integral component of *JMLA*. Each issue of *JMLA* typically includes between two and four Case Report articles. Topics featured with Case Reports vary, reflecting the breadth of services and initiatives that contemporary health science information professionals engage in across their local communities. Recent issues of *JMLA* have included descriptions of systematic review services, health information literacy programs, internship programs, and virtual conferences, among others.

Prospective *JMLA* authors often have trouble distinguishing whether a manuscript best fits within the Original Investigation or the Case Report category. Table 1 provides an overview of both submission types. While Original Investigations are slightly longer in extent, both submission types feature empirical articles that utilize structured abstracts, structured article formats, and are subject to *JMLA*'s Data Sharing Policy.

Table 1 A comparison of Original Investigations and Case Reports

	Original Investigations	**Case Reports**
Purpose	Describe research that employs any type of quantitative or qualitative method of analysis. Examples include intervention studies, surveys, content analyses, qualitative case studies, bibliographic or bibliometric analyses, and search filter development and testing.	Describe the development, implementation, and evaluation of a new service, program, or initiative, typically in a single institution or through a single collaborative effort.
Structured Abstract Format	Objective, Methods, Results, Conclusions	Background, Case Presentation, Conclusions
Structured Article Format	Introduction, Methods, Results, Discussion	Background, Case Presentation, Discussion
Extent	No more than 5,000 words; up to 6 figures and tables	No more than 3,000 words; up to 3 figures and tables
Data Sharing	Subject to the *JMLA* Data Sharing Policy	Subject to the *JMLA* Data Sharing Policy

DOI: dx.doi.org/10.5195/jmla.2025.2099

Given the substantive overlap between these two categories, we would like to highlight a few primary points of difference that delineate Original Investigations and Case Reports:

- *The inclusion of research questions:* Original Investigations are research projects launched to answer research questions; they will include research objectives and/or falsifiable hypotheses based on those questions. Case Reports typically describe local initiatives created based on the needs of a specific community; the premises guiding these reports are grounded in professional intuitions and assumptions.
- *A rigorous and well-defined research methodology:* Original Investigations must include a well-documented research methodology. While strong Case Reports include a program evaluation component, these evaluations are often focused on quality improvement rather than hypothesis testing, and may not have a rigorous methodology
- *A goal of generalizability*: Original Investigations attempt to answer broad research questions with results that have generalizable implications for the field. While strong Case Reports will share wider implications for other professionals to consider, Case Reports serve to bring to our attention to novel or surprising initiatives in local contexts that might not otherwise be described in the literature.

Authors sometimes express concerns over having their submissions placed within the Case Report category due to the misconception that Case Reports are a lesser publication type. While Original Investigation articles feature more rigorous research design and more intensive data analysis, Case Reports fill a critical role for *JMLA* readers: they present novel initiatives or provide preliminary findings that drive innovation and advance the practice of health information professionals. Moreover, prior to publication, Case Reports matriculate through the same rigorous double-blind peer review, editorial review, and copyediting processes as Original Investigations. In the remainder of this editorial, we will highlight the impact of recent *JMLA* Case Reports and discuss strategies authors can implement when drafting their own Case Report submissions.

THE IMPACT OF CASE REPORTS

Case Reports are highly valued by *JMLA*'s readership and widely read by library practitioners and information science researchers. Between September 2023 and September 2024, eight case reports received over 500 full-text views. Of the 100 most viewed *JMLA* articles during this period, 10 of these articles were Case Reports. Coincidentally, the most highly viewed *JMLA* Case Report during this period was Gotschall et al.'s "Journals Accepting Case Reports," which published a list of over 1,000 journals that publish case reports across dozens of medical specialties [3]. As with health information specialists, medical professionals value case reports as both an information source and a venue for disseminating their work.

The impact of Case Reports extends to citation practices, as well. We pulled citable *JMLA* articles (e.g., Knowledge Syntheses, Original Investigations, Case Reports, and Special Papers) published between 2019-2023 using the Web of Science Core Collection. While the most highly cited papers are knowledge synthesis studies, case reports perform similarly to original investigations in terms of citation impact. The median number of citations for all citable items is 3; the median for case reports is 2. Mean citations for all citable is 6.6 (SD 17.0); mean citations for case reports is 3.0 (SD 3.4). Of the top 40 articles in the last four years, four were published as Case Reports. Far from being an afterthought, well-written *JMLA* Case Reports on timely topics reach their intended audience and can shape professional practices.

ELEMENTS OF A STRONG CASE REPORT

Writing over 40 years ago to an audience of cardiologists, DeBakey and DeBakey [4] established several criteria for effective case reports that remain relevant for practitioners today. They contend that case reports should describe "unusual or puzzling features," depict "new, little known, or rare" occurrences, highlight "unexpected favorable or adverse" outcomes, or identify "possible causal relation, hitherto unreported, between two or more" items. While Case Reports within *JMLA* need not be entirely novel developments, the initiatives described should present a unique set of features, circumstances, or participants that separate them from previously published reports. As argued by DeBakey and DeBakey, valuable case reports "should uncover [a] truly unusual case from which others can learn something new"[4].

As such, Case Reports are not "light" or "easy" versions of Original Investigation articles, which seek to identify generalizable findings. Rather, Case Reports serve the distinctly different purpose of helping health science information professionals learn of surprising or innovative services or initiatives unfolding elsewhere within the field. To this end, Case Reports should describe the institutional setting, stakeholders, and other contextual information in sufficient enough detail for readers to understand the needs of the community from which the new initiative arose and consider whether the initiative could be equally beneficial within their local contexts. An effective Case Report also situates itself by mentioning some of the other possible solutions reported in the literature and making a case for why this novel approach improved upon these previously cataloged alternatives.

DOI: dx.doi.org/10.5195/jmla.2025.2099

While Case Reports might serve as the starting point for encouraging future generalizable original research studies, Case Reports need not feature the same in-depth data collection and analysis that is reported and discussed within Original Investigations. *JMLA* authors are encouraged to describe and report any relevant evaluation data that were gathered for the case. Example evaluation data to present within Case Reports may include attitudinal surveys, usage statistics, or responses from program participants. Inclusion of these data when available can enrich the Case Report, as these data can substantiate authors' claims about implications for professional practice while also establishing baseline findings to be further explored by readers.

However, some Case Reports suffer from paying too much attention to the evaluation process instead of describing the relevant context that made the case novel in the first place. *JMLA* often receives manuscripts that describe new services, programs, or initiatives whose evaluation data includes samples that are too small and non-representative to be meaningful, regardless of the robustness of the data collection and analysis methods used. In these instances, in-depth analysis of insufficiently powered studies may limit the authors' ability to adequately describe and reflect upon the service. While the implications do not have to be generalizable, strong case studies describe the authors' reflections on lessons learned.

In other instances, a Case Report's evaluation strategy may be sparse, but the program underlying the case is novel, important, and described objectively. Authors in these situations may benefit from describing the limitations of their evaluation process instead, rather than attempt to pull insights from such limited pools of data. Sparse data should not keep authors from considering the Case Report as a publication type for their innovative initiatives.

CONCLUSION

JMLA strives to enhance the knowledge base of health science information professionals through the publication of thoughtfully designed journal articles. *JMLA*'s Case Reports contribute to that mission by providing a forum for practicing health sciences librarians to highlight and share exciting programs occurring at their local contexts, regardless of whether these programs are research-based. The *JMLA* editorial team views Case Reports, when effectively written, as a valuable medium for driving innovation within professional practices. We hope this editorial will encourage new and previous *JMLA* authors to reflect on projects currently ongoing at the institutions and consider whether the programs' designs and outcomes may have the makings of a promising Case Report.

REFERENCES

1. Akers KG, Amos K. Publishing case studies in health sciences librarianship. *J Med Libr Assoc*. 2017;105(2):115-118. DOI: 10.5195/jmla.2017.212
2. Alpi KM, Evans JJ. Distinguishing case study as a research method from case reports as a publication type. *J Med Libr Assoc*. 2019;107(1):1-5. DOI: 10.5195/jmla.2019.615.
3. Gotschall T, Spencer A, Hoogland MA, Cortez E, Irish E. Journals accepting case reports. *J Med Libr Assoc*. 2023;111(4):819-822. DOI: 10.5195/jmla.2023.1747.
4. DeBakey L, DeBakey S. The case report. I. Guidelines for preparation. *Int J Cardiol*. 1983;4(3):357-364. DOI: 10.1016/0167-5273(83)90097-9.

AUTHORS' AFFILIATIONS

Jill T. Boruff, AHIP, jill.boruff@mcgill.ca, https://orcid.org/0000-0002-0338-7322, Co-Lead Editor, *Journal of the Medical Library Association*, Associate Librarian, Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University, Montreal, QC, Canada

Michelle Kraft, AHIP, FMLA, kraftm@ccf.org, Co-Lead Editor, *Journal of the Medical Library Association*, Medical Library Director, Cleveland Clinic Foundation, Cleveland, OH, United States

Alexander J. Carroll, AHIP, alexander.j.carroll@vanderbilt.edu, https://orcid.org/0000-0003-0248-3811, Associate Editor, *Journal of the Medical Library Association*, Associate Director, Stevenson Science and Engineering Library, Vanderbilt University, Nashville, TN, United States

Received October 2024; accepted October 2024

ISSN 1558-9439 (Online)

DOI: dx.doi.org/10.5195/jmla.2025.2089

Amy Blevins, Medical Library Association President, 2023-2024

Elizabeth Kiscaden; Hannah J. Craven; Gabriel R. Rios; Ryan Harris; Joey Nicholson

See end of article for authors' affiliations.

Amy Blevins served as the Medical Library Association president from 2023-2024. In this presidential biography, the authors outline a history of Blevins' recruitment to the career, career development, and impact on the association and the profession.

Amy Blevins' colleagues, mentors, and friends always knew it was not a matter of if, but rather when she would serve as the president of the Medical Library Association. Enthusiastic, driven, highly intelligent, and deeply engaged in the profession, Blevins stands apart for her passion for health sciences librarianship and commitment to advancing the profession. What is more, throughout the highs and lows of building our association or progressing in her own career, Blevins makes it fun. Have you ever served on a committee with Blevins? If you did, you would remember it, as she likely had you laughing!

Don't let Blevins' animated nature or unconventional interests lead you to underestimate her professional career. In your acquaintance with Blevins, she may have educated you about the harmless, yet slimy, hellbender salamander, which is a threatened aquatic species. Or, given her love for horror movies, she may have shared the knowledge that the movie Critters was based on a nightmare the producer had when he was a child. While discussing the cryptids of North America or admiring her signature goth earrings, do not forget that Blevins is a research powerhouse and one of the most skilled instructors of evidence based practice in our field. Blevins has a commitment to staying true to who she is and demonstrates how to be a professional at the top of the field on her own terms.

RECRUITMENT TO THE PROFESSION

Blevins began her undergraduate journey at Saint Louis University where she pursued a Bachelor of Arts in Biology. She was already familiar with the university; her mother had gone back to school when Blevins was a teenager; given that she was a single mother, she brought Blevins along to study in the library. As an incoming college student, Blevins applied for a job in the library, but was soon overachieving, early evidence of the internal drive that continues to motivate her throughout her career. It wasn't long before Blevins was promoted to an assistant role, supporting the reference services staff in the library.

Blevins' career plan was to continue on to graduate school to earn an MD/PhD in order to work as a genetic researcher, an idea inspired by a middle school program designed for talented students. It was thanks to the influence of a number of outstanding librarians she crossed paths with at the library that she changed course. She had many great supervisors and mentors at the Pius XII Memorial Library including (but not limited to): Jeannette Pierce, Martha Allen, Jamie Emery, Jonathan Harms, Phill Barron, and Claudia DuVall. Blevins came to find the academic library a second home and found a

DOI: dx.doi.org/10.5195/jmla.2025.2089

strong sense of belonging among the academic librarians who worked there-she had found her people.

Blevins went on to complete her Master of Arts in Library Science at the University of Missouri - Columbia, where she continued to work in the academic library. In fact, this is where she first gained experience working in a health sciences library, where she completed a tenure as a graduate library assistant at the J. Otto Lottes Health Sciences Library. Although she originally intended to become a science librarian, her experience as a graduate library assistant solidified her intent to specialize in health sciences librarianship. It was at J. Otto Lottes that Blevins found her calling, and thanks to supportive mentors, such as Rebecca Graves, Amanda McConnell, and Diane Johnson, that she found a passion for Medical Subject Headings (MeSH) and advanced search strategies in biomedical databases.

EAST CAROLINA UNIVERSITY

As a newly minted graduate, Blevins accepted a position as a librarian liaison to the College of Allied Health Sciences and College of Health and Human Performance at East Carolina University (ECU) in 2006. The role matched her interest in instruction and she set her sights on developing online instructional materials at an institution that prided itself on being the largest provider of distance education in the state at that time. During her tenure at East Carolina University, Blevins partnered on the development of a new web page to display electronic resources, led an evaluation of online course content, and conducted an evaluation of software tools used for developing content.

It was through Blevins' work designing and delivering online instruction and creating tutorials that she moved into the newly created position of Education and Instructional Technology Librarian at ECU. Developing herself further in this role, Blevins designed longitudinal curriculum-based instruction for an occupational therapy program, which had an evidence-based practice component. She went on to lead the Tea Time Training program for training staff internally and to host a podcast with the main academic library titled, *Research First Aid*.

It is worth noting that Blevins' research posters from this chapter of her life feature outstanding original art. At East Carolina University, Blevins partnered with graphic designer, Jason Cottle, who went on to become a lifelong personal friend. Due to this connection, research posters from this time period feature animated scenes, such as a kitchen in which the co-presenters are featured chopping up the "ingredients" for information literacy skills (see Figure 1). The "active learning elements" are ready to pitch into the cooking pot, and from within that pot, a long, pink squid tentacle can be seen, draping over the edge.

Blevins often acknowledges the William E. Laupus Library at East Carolina University for supporting her early engagement in the profession regionally and nationally. The library provided professional development funds which allowed Blevins to attend the Medical Library Association's Mid-Atlantic Chapter (MAC) meetings, where she quickly got involved in the organization and made lifelong connections. It was her engagement in MAC that connected her with Shannon Jones, another former MLA president and a legend within the profession of health sciences librarianship. Jones was Blevins' conference mentor at her first health sciences librarian conference and has mentored her throughout her career.

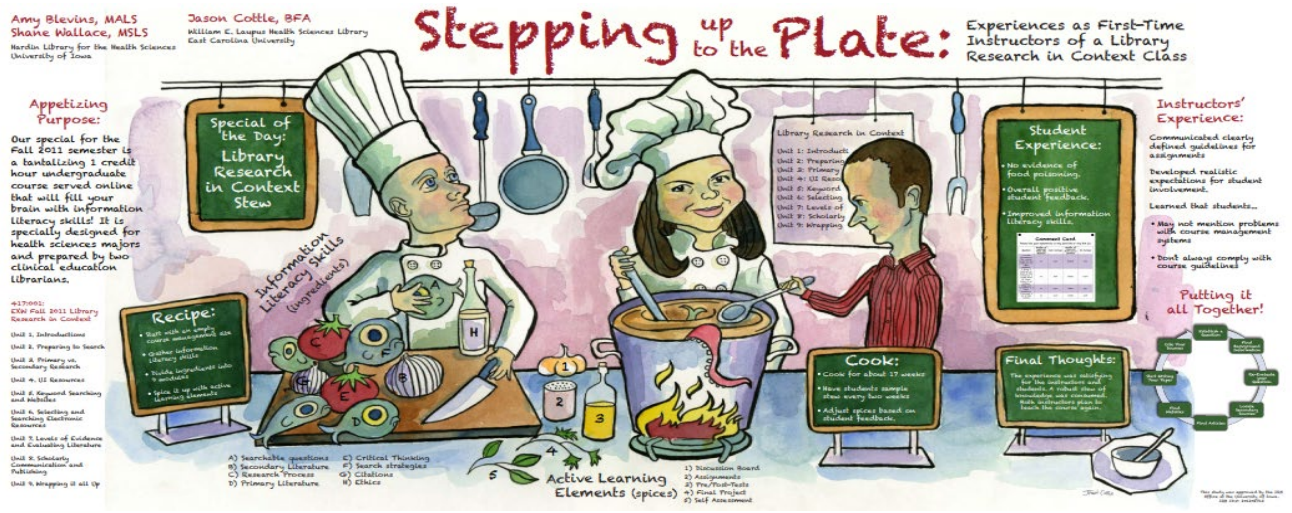


Figure 1 Amy Blevins, & S. Wallace. "Stepping Up to the Plate: Experiences as First-Time Instructors of a Library Research in Context Class." Medical Library Association Annual Conference, 21 May 2012.

DOI: dx.doi.org/10.5195/jmla.2025.2089

UNIVERSITY OF IOWA

Seeking to further develop herself, Amy Blevins accepted a Clinical Education Librarian position at Hardin Library for the Health Sciences at the University of Iowa. Within this role, she served as the liaison to the School of Medicine and received a faculty appointment within the Carver College of Medicine. Blevins wasted no time in establishing herself in the new role, catalyzing her colleagues and expanding her professional network throughout the upper Midwest.

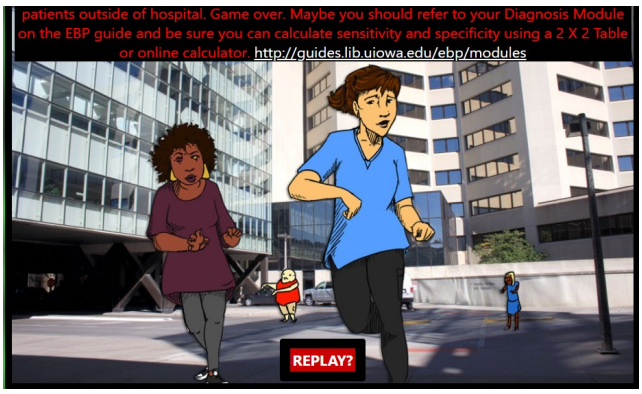
As a part of her role, Blevins was responsible for managing the Simulation Center, which was housed within and managed by the health sciences library. Under her leadership, Blevins collaborated with colleagues to expand a program which used simulation equipment as a tool for outreach to area school-aged children. Children from grades 2 to 12 would visit the Simulation Center to try tying surgical knots, use the eye simulators, use three-dimensional anatomy tools, and try out "Harvey," the cardiac rhythm simulator located within the Center.

Another job responsibility assigned to Blevins was utilizing technology to improve library services. These were relatively early days for online library instruction; Blevins built on her prior experience to collaborate with colleagues on the development of best practices, policies, and procedures for the creation of online instruction. With a team, Blevins evaluated software tools to use, developed standard opening and closing screens for tutorials, and presented training sessions to library staff interested in creating their own content.

In 2011, Hardin Library's director, Linda Walton, talked about a new potential service for supporting systematic reviews. After attending the University of Pittsburgh workshop, Blevins started supporting researchers and residents. As demand for the service grew, Blevins worked with her colleagues to launch a formal systematic review service at Hardin Library. After seeking solutions from colleagues on Medlib-L (the Medical Library Association's listserv), she developed a Memorandum of Understanding (MOU), at a time when there were few examples available. She shared her library's work with others and MOUs continue to be used to ensure librarians' contributions are recognized through authorship.

Back when the zombie television show, *Walking Dead*, was popular, Blevins partnered with her colleague (and co-author of this manuscript), Elizabeth Kiscaden to develop and pilot a computer game designed to teach medical students critical appraisal skills. Combining Blevins' love of zombies with Kiscaden's love of 90s-era adventure computer games, the team used an internal grant to launch a story based apocalypse game. The game lives on (https://www.lib.uiowa.edu/hardin/zombies-ate-my-evidence/), with timeless art featuring zombies in multiple locations across the University of Iowa Health Center (see Figure 2).

Figure 2 Zombies Ate My Evidence. 2017.



INDIANA UNIVERSITY SCHOOL OF MEDICINE

The next step in Blevins' career led her to her current role as Associate Director for Public Services at the Ruth Lilly Medical Library (RLML) at the Indiana University School of Medicine (IUSM). Since 2015, Blevins has made significant strides in aligning the library's public services, education, and outreach programs with the school's mission, including leading an evidence-based medicine (EBM) thread within the School of Medicine and developing a multi-tiered systematic review service.

One of Blevins' most significant contributions in this area is her work designing and implementing EBM instruction for medical students. Recognizing the importance of integrating EBM skills throughout the medical curriculum, Blevins advocated for, and created a scaffolded approach to teaching these skills. This method, inspired by constructivist learning theory, ensures that students build upon their knowledge and abilities as they progress through their medical education, preparing them to be effective, evidence-based practitioners.

Typically, the leadership of major curricular components in medical schools, such as EBM education, is entrusted to clinicians who have hands-on experience in applying these principles in patient care. However, Blevins was appointed as the leader of the EBM thread at IUSM. This appointment represents a significant departure from the norm and this exceptional recognition of Blevins' expertise in teaching EBM forages the unique and valuable contributions that librarians can make when integrated into medical education courses.

As the EBM thread leader, Blevins took on the crucial responsibility of guiding an interdisciplinary team of clinicians, biostatisticians, and other healthcare professionals, each bringing their unique perspectives and expertise to the table. Under her leadership, they developed a comprehensive scaffolded EBM curriculum.

DOI: dx.doi.org/10.5195/jmla.2025.2089

These objectives are designed to ensure that students acquire the necessary skills to practice evidence-based medicine effectively.

Another significant contribution to the Ruth Lilly Medical Library (RLML) and the broader research community at Indiana University School of Medicine has been her creation and leadership of a formalized Systematic Review service. Recognizing the growing importance of systematic reviews in evidence-based healthcare and building on her experiences and Iowa's Hardin Library, Blevins took the initiative to create and lead a taskforce dedicated to formalizing the library's systematic review services. This move was particularly crucial given the time-intensive nature of systematic reviews, which typically take 12-18 months to produce, and their status as the highest level of research evidence when conducted properly.

Blevins' approach to the systematic review service emphasized the valuable role that librarians play in these high-impact publications. She ensured that librarians were positioned as integral team members, responsible for designing and executing comprehensive literature searches, writing the search methods sections of final publications, and advising research teams on the standards for conducting and reporting systematic reviews. The impact of this initiative has been significant as this marked a notable shift from previous practices, where librarians were more likely to be mentioned in the acknowledgments section rather than recognized as full collaborators.

The importance of this work extends beyond the library and even the university. By ensuring that RLML librarians are providing information to research teams about best practices for systematic reviews, Blevins is contributing to the publication of high-quality research that directly supports patient care. This aligns perfectly with the broader mission of evidence-based healthcare and underscores the critical role that skilled librarians play in the research ecosystem.

NATIONAL LEADER IN EVIDENCE-BASED PRACTICE

Amy Blevins' evolution into a national leader in Evidence-Based Practice (EBP) is a testament to her dedication to advancing the field of health sciences librarianship. Her involvement in key initiatives and workshops has solidified her reputation as an expert in this crucial area.

Blevins' local work at RLML set the foundation for her involvement and leadership with the Critical Appraisal Institute for Librarians (CAIFL). Blevins was recruited by Marie Ascher in 2018 to serve on a steering committee to develop CAIFL, and the first cohort completed the program in the spring of 2019. Blevins serves as a small group facilitator for CAIFL, guiding participants through the intricacies of critical appraisal techniques. This role allowed her to share her expertise while also learning from the diverse experiences of participants from across the country. The program has been a success in developing the next generation of EBP leaders in health sciences librarianship and plans are underway to offer the institute again in 2025.

Another significant milestone in Blevins' development as a national EBP leader was her involvement in the Evidence-Based Practice for Health Sciences Librarians Workshop (EBP for HSLs). Blevins' involvement began with her being invited to serve as a teaching fellow in 2018 and has transitioned into a leadership role for the national workshop. For the reimagined 2023 workshop, Blevins collaborated with a co-facilitator and teaching fellow to create materials for both large and small group sessions. Blevins independently created a new introduction to critical appraisal, including an innovative video component accompanied by a quiz for pre-work.

The impact of Blevins' work in these national initiatives extends beyond the workshops themselves. The materials and approaches developed for these programs influence local practices at institutions across the country.

Blevins' work in these national initiatives extends beyond the workshops themselves, the content developed influences local practices at institutions across the country. Blevins' work not only advances the field but also sets a standard for how EBP can be taught and implemented in library and medical education settings. As she continues to contribute to these national initiatives, Blevins is shaping the future of EBP in health sciences librarianship, ensuring that librarians remain at the forefront of evidence-based healthcare education and practice.

LEADERSHIP IN PROFESSIONAL ASSOCIATIONS

Blevins' career is marked by a strong commitment to professional service and leadership. Her dedication to service started early on in her career when she served as chair of the MAC's Membership and Recruitment Committee. This was just the beginning of her varied and wide-ranging service to medical librarianship. Her dedication to the Medical Library Association (MLA) stands out as particularly impactful. Blevins' most recent service for the Medical Library Association (MLA) was as MLA President from 2022-2025. Blevins also served as a member of the executive board and treasurer from 2016 to 2019, Blevins played a crucial role during a period of significant change for the organization. Her leadership in financial matters, including the creation of a new Finance Committee as the MLA Treasurer, helped steer the MLA through this transitional period.

Her leadership extends to various sections within the MLA, including her work with the Educational Media and Technologies Section (EMTS), for which she served in multiple leadership roles including as chair. Under her

DOI: dx.doi.org/10.5195/jmla.2025.2089

guidance, EMTS developed free continuing education opportunities for members, an initiative that was recognized with the Section Project of the Year award. This project exemplifies Blevins' commitment to making professional development accessible to all librarians in her field.

Blevins' work for the profession has been recognized in several different ways including being awarded the Lucretia W. McClure Award for Teaching Excellence and with several colleagues, the Ida and George Eliot Prize. While Blevins has served in many capacities for MLA, she always takes the time to foster the growth of new health sciences librarians. Blevins consistently serves as a mentor for first time conference attendees through MLA's Colleague Connection program.

Blevins' leadership abilities have been recognized beyond her immediate professional circles. Her selection for the NLM/AAHSL Leadership Fellows Program, a highly competitive program designed to prepare librarians for director positions, speaks to her potential for top-level leadership in academic health sciences libraries. At the program's capstone Amy was selected to give the keynote talk on behalf of the fellows. This experience has undoubtedly shaped her approach to management and strategic planning in her current role.

LOOKING BACK, LOOKING FORWARD

Blevins' impact on the profession includes over 45 presentations, 30 journal articles, and four books or book chapters - evidence of a highly productive professional career. Beyond these scholarly contributions, Blevins has developed and presented more than 25 courses on topics ranging from evidence based practice, literature searching for systematic reviews, online instruction, assessing information needs, and critical appraisal. She has served on countless committees within universities and has been active within the Mid-Atlantic and Midwest Chapters of the Medical Library Association, as well as being active at the national level, culminating in her tenure as MLA president.

Something particularly notable about Blevins' scholarly productivity is her practice of continuously inviting colleagues to partner with her on projects. These invitations have provided early career librarians she mentors with an opportunity to build their CVs and brought some of her former mentors back to collaborate on scholarship. Those who have worked alongside Blevins continue to be invited back for partnership on new scholarly endeavors.

Considering that Blevins is still mid-career, one wonders what future contributions she will make to the health sciences profession. While colleagues have tried to attract her to roles in the larger sphere of academic librarianship, she remains persistently committed to the discipline of health sciences. It was in this discipline that Blevins found belonging and has been only too willing to give back to the profession.

Wherever Blevins invests the second half of her career, one can count on her contributing bold ideas, putting in the work to realize change, collaborating with colleagues, and making the entire exercise fun. In the process of interviewing Blevins' former mentors for this article, authors asked them to share any final memories. In several cases, interviewees stated the same thing, "I can think of several funny stories… but they are not appropriate for the article!" But all also shared a sentiment similar to this comment from Lisa Traditi, "Amy lightens the room when she's in it, being around Amy makes you feel like everything is going to be okay even when it's hard." Given the challenges within academic librarianship and higher education, couldn't we all use more leaders like that?

ACKNOWLEDGMENTS

The authors would like to acknowledge the following individuals: Kaushar Mahetaji (University of Toronto), who assisted with the survey implementation; Heather Ganshorn (University of Calgary), Julie Marcoux (Dalhousie University), Judith Logan (University of Toronto) and Heather Cunningham (University of Toronto), who provided valuable feedback on the draft survey questions.

AUTHORS' AFFILIATIONS

Elizabeth Kiscaden, elizabeth.kiscaden@uc.edu, https://orcid.org/0000-0001-7230-030X, University of Cincinnati, Cincinnati, OH

Hannah J. Craven, hancrave@iu.edu, https://orcid.org/0000-0002-1701-3655, Ruth Lilly Medical Library, Indiana University School of Medicine, Indianapolis, IN

Gabriel R. Rios, grrios@iu.edu, https://orcid.org/0000-0002-4479-8670, Ruth Lilly Medical Library, Indiana University School of Medicine, Indianapolis, IN

Ryan Harris, rharr103@charlotte.edu, https://orcid.org/0000-0003-3322-1371, UNC Charlotte J Murrey Atkins Library, University of North Carolina, Charlotte, NC

Joey Nicholson, Joey.Nicholson@nyulangone.org, https://orcid.org/0000-0001-8658-5879, NYU Langone Health, New York, NY

Received October 2024; accepted December 2024

ISSN 1558-9439 (Online)

KNOWLEDGE SYNTHESIS
DOI: dx.doi.org/10.5195/jmla.2025.1965

A scoping review of librarian involvement in competency-based medical education

John W. Cyrus; Laura Zeigen; Molly Knapp; Amy E. Blevins; Brandon Patterson

See end of article for authors' affiliations.

Objective: A scoping review was undertaken to understand the extent of literature on librarian involvement in competency-based medical education (CBME).

Methods: We followed Joanna Briggs Institute methodology and PRISMA-ScR reporting guidelines. A search of peer-reviewed literature was conducted on December 31, 2022, in Medline, Embase, ERIC, CINAHL Complete, SCOPUS, LISS, LLIS, and LISTA. Studies were included if they described librarian involvement in the planning, delivery, or assessment of CBME in an LCME-accredited medical school and were published in English. Outcomes included characteristics of the inventions (duration, librarian role, content covered) and of the outcomes and measures (level on Kirkpatrick Model of Training Evaluation, direction of findings, measure used).

Results: Fifty studies were included of 11,051 screened: 46 empirical studies or program evaluations and four literature reviews. Studies were published in eight journals with two-thirds published after 2010. Duration of the intervention ranged from 30 minutes to a semester long. Librarians served as collaborators, leaders, curriculum designers, and evaluators. Studies primarily covered asking clinical questions and finding information and most often assessed reaction or learning outcomes.

Conclusions: A solid base of literature on librarian involvement in CBME exists; however, few studies measure user behavior or use validated outcomes measures. When librarians are communicating their value to stakeholders, having evidence for the contributions of librarians is essential. Existing publications may not capture the extent of work done in this area. Additional research is needed to quantify the impact of librarian involvement in competency-based medical education.

Keywords: Competency-Based Education; CBME; Evidence-Based Medicine; EBM; Problem-based learning; case-based learning; entrustable professional activities; self-regulated learning; lifelong learning; librarians; libraries; Instruction; education; learning; curriculum; training; undergraduate medical education

See end of article for supplemental content.

INTRODUCTION

The Association of Academic Health Science Libraries (AAHSL) formed the Competency-based Medical Education (CBME) Task Force on March 3, 2016, in order to identify the nature and depth of AAHSL Libraries' participation in the changes taking place in medical curricula highlighted by the adoption of Core Entrustable Professional Activities (Core EPAs). Competency-based medical education "is an outcomes-based approach to the design, implementation, and evaluation of education programs and to the assessment of learners across the continuum that uses competencies or observable abilities"[1]. Core EPAs, published in 2014 by the Association of American Medical Colleges (AAMC), provide a structure by which to measure the 13 basic competencies required by the Accreditation Council for Graduate Medical Education (ACGME) for residents going into their first day of residency. The EPAs represent the skills residents should learn in their undergraduate medical education. In particular, the EPAs include developing a well-formed clinical question to retrieve evidence to support clinical decision-making (EPA 7) and collaborating as part of an interprofessional team (EPA 9) [2]. Both of these are areas in which librarians have a vested interest and participate in the educational process of moving learners from pre-entrustable to entrustable. Thus, it is of critical importance to librarians working in medical education to understand where these competencies are being assessed. The work of the original AAHSL CBME task force resulted in the mapping of EPAs

DOI: dx.doi.org/10.5195/jmla.2025.1965

to the Association of College Research Libraries (ACRL) Information Literacy Framework and a survey of librarian involvement with EPA 7 that was later published in BMC Medical Education [3]. In August of 2019, a new AAHSL taskforce was established and charged with implementing competency-based medical education (CBME) taskforce recommendations. One of the goals of the new taskforce was to "create a collection of case studies, vignettes, best practice stories, or other representations demonstrating the beneficial roles and positive impacts of librarian engagement in competency-based medical education (CBME)."

The new task force referred to the work of the previous task force and examined relevant literature to guide their work. In their 2012 review, Dorsch and Perry found that while there were numerous studies published on the topic of librarian involvement in medical education, "gaps in the literature suggest a need for longitudinal follow-up and multicentered studies to validate the findings of the literature to date"[4]. A scoping review was selected for this research as the methodology lends itself both to the mapping of an area of research and the identification of gaps in existing research [5]. This scoping review seeks to understand the current state of librarian involvement in CBME and provide demonstrable evidence of the value of engaging in this work to both librarians and medical education stakeholders. Specifically, the review sought to answer what roles librarians play in supporting CBME, how interventions involving librarians are designed, which outcomes have been used to measure the impact of librarian work in CBME, and whether or not there is evidence that any of these outcomes affect clinical competence?

METHODS

We performed a scoping review of published literature on librarian involvement in competency-based medical education in accordance with guidance from the Joanna Briggs Institute (JBI) Manual for Evidence Synthesis [6] and reported following the PRISM-ScR guidelines [7]. The protocol for this review is available through the Open Science Framework (https://osf.io/gcy4e).

The authors used the Association of American Medical Colleges' definition of CBME as "an outcomes-based approach to the design, implementation, and evaluation of education programs and to the assessment of learners across the continuum that uses competencies or observable abilities"[8]. In order to operationalize this definition for this review, the following concepts were included to describe content falling under the umbrella of CBME: entrustable professional activities (EPAs), self-directed learning (SDL), evidence-based medicine (EBM), interprofessional education (IPE), quality improvement, systems-based practice, health systems science, health services research, translational science, shared decision making, case-based learning, and problem-based learning.

The research team conducted searches in the following databases: Medline (Ovid), Embase (Ovid), ERIC (EBSCO), CINAHL Complete (EBSCO), SCOPUS (Elsevier), and Library & Information Science Source (LISS)/Library Literature & Information Science (LLIS)/Library, Information Science & Technology Abstracts (LISTA) via EBSCO. No multi-database searching was conducted. Each database was searched individually. An initial search was run on April 14, 2021, and an updated search was run to include articles published up to December 31, 2022. To be inclusive, controlled vocabulary terms and keywords for the concepts of competency-based medical education, critical thinking, evidence-based practice, and libraries/librarians were used. The concept "libraries/librarians" was specifically added since, without this, the search might return a body of results comprised of all the literature about CBME, not just the subsection mentioning librarians and libraries in the context of CBME. No filters for study type, date, or language were used. The search results were imported into Covidence systematic review management software (https://www.covidence.org/). Duplicate records were removed using Covidence. Full search strategies are included as Supplementary Material.

All screening took place in Covidence in two phases: title/abstract and full-text. Selection was conducted independently with two reviewers screening each study. Conflicts were resolved by consensus among the entire team. Eligibility criteria were established a priori. To be included in the review, papers had to describe librarian involvement in the planning, delivery, and/or assessment of competency-based medical instruction or educational intervention in undergraduate medical education (UME), the phase of medical education that confers the Doctor of Medicine (MD) degree. Additionally, studies need to be conducted in Liaison Committee on Medical Education (LCME) accredited medical schools located in the United States. The LCME is the accrediting body for education programs in the United States leading to an MD degree. Studies that were not published or available in the English language were excluded.

Following the process for charting described in Arksey & O'Malley [5], we extracted the following variables from each study into a spreadsheet generated using Google Forms: author name, date of publication, the title of the journal, the competency domain(s) assessed (based on EPAs where librarians self-identified involvement), and whether or not the outcomes addressed clinical competence. The components of EPA 7, which includes elements of EBM, were further mapped to four of the five A's of the EBM cycle. The competencies we assessed included:

DOI: dx.doi.org/10.5195/jmla.2025.1965

- EPA 7 - Ask: Developing a well-formed, focused clinical question
- EPA 7 - Acquire: Awareness and skills in using information technology to access accurate and reliable medical information
- EPA 7 - Appraise: Skills in appraising sources, content, and applicability of evidence
- EPA 7 - Apply: Apply findings to individuals or populations, communicate findings to patient and team, reflecting on process
- EPA 9 - Identify team member roles and responsibilities and seek help other members of the team to optimize health care
- EPA 9 - Include team members, listen attentively, adjust communication content and style to align with team-member needs
- EPA 9 - Establish and maintain a climate of mutual respect, prioritize team needs over personal needs.

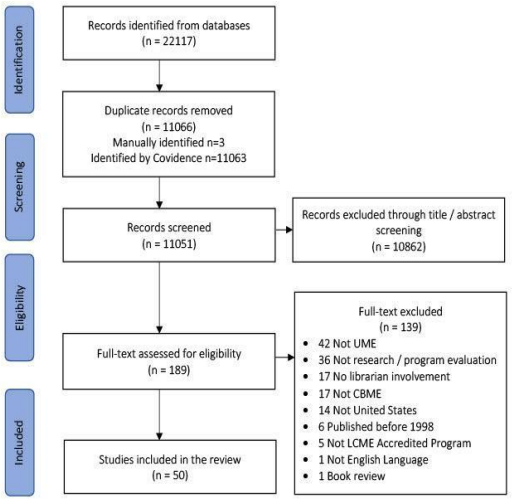
For empirical studies, defined as quantitative studies for this review, we extracted the dates of data collection, study aim, location of research, name of institution where research was conducted, population, intervention/exposure, duration of intervention/exposure, and librarian role in curriculum. The librarian roles in the curriculum were defined by the authors as follows based on the synthesis of existing literature: collaborator (librarian is not the instigator but involved in the teaching), curriculum designer (primarily involved in designing the curriculum), leader (instigator of curriculum or session), or evaluator (directly involved in the evaluation of student skills and knowledge gained through the curriculum) [4, 9, 10, 11]. The purpose of the study (program/curriculum evaluation, course/class evaluation, program/curriculum/course development, curriculum review/mapping), study design, direction of findings by outcome (positive, no change, negative, not reported), and the measure used for outcomes assessment were also extracted. Study outcomes were categorized by Kirkpatrick Model level [12]. The levels of this model, which is used to conceptualize how training is evaluated, includes reaction (learner reaction to and thoughts about their training experience), learning (learner change in knowledge from baseline as a result of the training), behavior (observable, measurable, repeatable behavior that the learner can demonstrate), and results (the tangible results of the training, such as improved patient outcomes). For evidence synthesis studies, which included both narrative reviews and more formal methodologies like systematic reviews, we collected the study aim/question(s), population/setting of interest, number and names of databases searched, date of last search, review design (literature, systematic, meta-analysis, scoping review, etc.), number of studies included, and the findings related to aim/research question (positive, no change, negative). The data extraction form was piloted with the entire group. Two reviewers extracted data from each study with a third reviewer to check the data and resolve conflicts.

We used descriptive statistics to describe the extent, nature, and distribution of the studies included in the review. In addition, we analyzed data related to publication dates and journals for all included studies. Studies were categorized by the characteristics of the interventions and by the levels of outcomes and how the outcomes were measured. Risk of bias assessment was not conducted for this scoping review as it was deemed not to provide useful information relevant to the research questions addressed by this review.

RESULTS

Of the 11,051 studies screened for inclusion, 50 were included (Figure 1 PRISMA Flow Diagram). Forty-six articles were empirical research or program evaluation and four were some form of evidence synthesis. The articles were published in eight journals, including Academic Medicine (5), BMC Medical Education (1), BMJ Evidence Based Medicine (2), Health Libraries Review (1), Journal of the American Medical Informatics Association (1), Journal of the Medical Library Association or the Bulletin of the Medical Library Association (15), MedEdPORTAL (1), and Medical Reference Services Quarterly (24) between 1996 and 2022.

Figure 1 PRISMA Flow Diagram



DOI: dx.doi.org/10.5195/jmla.2025.1965

Characteristics of the Interventions

The teaching strategies employed as part of the intervention varied, and generally included didactic lecture followed by individual or small group work. Content and structure were inconsistent. All teaching strategies reported positive outcomes. Teaching strategies included in person didactic lecture, online learning modules, pre-recorded lectures, PubMed practice searches, clinical case worksheets, presentations, small group work, and problem-based learning (see Table 1 Characteristics of interventions from reviewed studies).

The duration of the intervention also varied. On one end there was a 30-minute lecture on PubMed/MEDLINE and on the other end was teaching concepts longitudinally throughout an entire semester or over several years. While formal statistical analysis was not conducted to test the relationship, the duration of the intervention did not appear to correlate with positive results. The studies without specifically reported positive results included a semester-long class and multiple 60 to 120-minute one-shot interventions. Librarians were active in every role that we used as a category with the most common role being Collaborator (36) followed by Curriculum Designer (30), Leader (16), and Evaluator (10) (see Table 2 Librarian roles from review studies).

The most common domains covered by interventions were EPA 7 Appraise (n=39, 78%), followed by EPA 7 Ask (n=32, 64%) and EPA 7 Acquire (n=29, 58%). Other domains were covered less extensively, including EPA 7 Advise (n=15, 30%) and all domains associated with EPA 9 (Team Roles (n=2, 4%), Mutual Respect (n=1, 2%), and Team Communication (n=1, 2%).

The majority of studies used unvalidated outcomes measures and there was little consistency among the outcomes that were assessed (see Table 3 Outcomes of interventions from reviewed Studies). Of the tools used to measure the effect of the interventions, three studies utilized a modified Fresno test and one used the Berlin questionnaire [15, 28, 40]. The Fresno test and Berlin questionnaire are two of a handful small number of validated scales that assess competence in evidence-based practice knowledge and skills [58, 59]. Seven studies used a pre-/post-intervention evaluation design, largely through anonymous/online surveys [23, 24, 33, 34, 38, 47, 50]. The remaining studies required students to synthesize or actively apply knowledge asking them to develop a case scenario and make a case or team presentation, perform in an objective structured clinical examination (OSCE) case study, create a patient-centered disease information resource sheet, answer questions that asked them to identify the highest quality of evidence in sources, or submit multiple MEDLINE search strategies that were then evaluated by librarians. Three studies used a rubric to evaluate assignments, but details on the composition or creation of the rubric were not given [31, 48]. Only two studies specifically mentioned giving formative feedback to learners [31, 36].

Study outcomes were categorized by Kirkpatrick Model level, which describes outcomes by the type of data that they collect and what that data conveys. The majority of studies assessed satisfaction and knowledge retention outcomes with 30 (65%) looking at reaction and 26 (57%) looking at knowledge outcomes. Fewer studies looked at outcomes that might transfer to clinical practice, such as the impact of an intervention on behaviors or how the interventions impact downstream results (learner, patient, clinical outcomes) with 11 (24%) looking at behavior and 6 (13%) looking at results as outcomes. More than 95% of studies reported positive outcomes; however, no study directly addressed the clinical competence of the learners.

Characteristics of Evidence Syntheses

Four evidence syntheses articles explored various ways librarians involve themselves in CBME [4, 10-11, 60]. Out of 17 databases, the most commonly searched databases regardless of platform were MEDLINE (n=4), CINAHL (n=3), Embase (n=2), Web of Science (n=2), Scopus (n=2), ERIC (n=2), and PsycINFO (n=2). One article was a narrative review that reported methods but did not adhere to a specific methodological framework [4]. The remaining articles following scoping review [60] and systematic review methodologies [10,11]. All evidence synthesis papers aimed to describe and assess instructional methods for teaching evidence-based practice concepts and skills, including searching, to health sciences or medical students. All syntheses reported results that trended positive but varied significantly from study to study. All studies also reported challenges in synthesizing evidence based on the diversity of interventions and outcomes measures, and a lack of standardized assessment tools. These studies also highlighted the disparate roles played by librarians in instruction, ranging from lecturer to curriculum designer, and the need to report detailed, standardized descriptions of educational interventions.

DOI: dx.doi.org/10.5195/jmla.2025.1965

Table 1 Characteristics of interventions from reviewed studies

Study ID	Population	Intervention	Duration of Intervention	Librarian Role (Leader, Collaborator, Curriculum Designer, Evaluator)	Domains Covered by Intervention*
Abate et al., 2011 [13]	1st year medical students	Four didactic sessions on resources, searching, and evidence-based medicine.	90 minutes x 4	Leader, Curriculum Designer	EPA 7-Acquire
Adams, 2015 [14]	1st year medical students	Course	Two weeks	All of the above	EPA 7-Ask
Aronoff et al., 2017 [15]	Students from nine health professions across two institutions (medical, dental, pharmacy, nursing, occupational therapy, physical therapy, social work, speech language path, dietetics)	Two online learning modules ("Intro to EBP" and "Finding Evidence in PubMed") followed by facilitated in-person small group case-based learning experience.	Two hours (1 hour per module)	Collaborator, Curriculum Designer, Evaluator	EPA 7-Acquire
Blake et al., 2018 [16]	1st and 2nd year medical students	Pre-recorded evidence-based medicine modules (Interviewing a standardized patient, Practice reaching a diagnosis, Practice searching PubMed and point of care tools for evidence)	A single 4-hour in-person session plus 3 hours of pre-recorded lectures	Collaborator	EPA 7-Ask
Blanco et al., 2014 [17]	Deans from AAMC medical schools	A cross-sectional survey.	n/a	Collaborator	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise
Blumenthal et al., 2005 [18]	1st year medical students and 3rd year medical students on their family medicine rotation.	1st Year Students: large group sessions led by faculty-librarian team with student presentations, 3rd Year: 2-hour small group led by faculty-librarian team with students completing an EBM clinical case worksheet.	1st Year: not reported, 3rd Year: 2 hours	Collaborator	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply
Brahmi et al., 1999 [19]	4th year medical students	Five two-hour sessions taught across one week (2 hours each day) on EBM, searching MEDLINE and Cochrane, and critical appraisal of research.	1 week: 2 hours a day for 5 days	Collaborator	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply
Brown and Nelson, 2003 [20]	1st and 2nd year medical students	Longitudinal instruction in constructing clinical questions, searching skills, and library resources.	Multiple sessions over two years	Collaborator, Evaluator	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply
Burrows and Tylman, 1999 [21]	3rd year medical students	Evaluation of literature searches before and after information resources and searching skills training.	1996-1998	Evaluator	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise

DOI: dx.doi.org/10.5195/jmla.2025.1965

Butera et al., 2014 [22]	1st year medical students	A combination of case-based scenarios and web-based information resources tailored to the assignment with direct librarian support for student research questions embedded into the course.	Semester length course	Collaborator, Curriculum Designer	EPA 7-Ask, EPA 7-Acquire
Cyrus et al., 2013 [23]	3rd and 4th year medical students	Described as a "selective" comprising two sessions: a library session on database searching and statistical concepts, a session on critical appraisal of preselected articles to emphasize statistics and research design, and a session on critical appraisal of articles submitted by students and re-emphasis of statistical concepts.	2 or 3 sessions	Collaborator	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise
Dorsch et al., 2004 [24]	3rd year medical students	An evidence-based medicine seminar series of in-person group sessions	8 one-hour seminars during a 12-week internal medicine clinical rotation.	Collaborator, Evaluator	EPA 9 - Identify
Earl, 1996 [25]	1st year medical students	A problem-based learning case and group work.	One-hour group work during class time.	All of the above	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise
Eldredge et al., 1998 [26]	Librarians	Program description of a School of Medicine curriculum reform.	Not reported	Collaborator, Curriculum Designer	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply
Eldredge et al., 2021 [27]	1st year medical students	A pre-post-design with the intervention consisting of a lecture on question formulation, a case vignette, and practice formulating clinical questions from the vignette.	Single session (duration not reported), including a 25-minute lecture from a librarian.	All of the above	EPA 7-Ask, EPA 7-Acquire
Gagliardi et al., 2012 [28]	3rd year medical students	A combination of large group lecture and case-based learning team taught by librarians and diverse clinical faculty.	Six two-hour sessions over six consecutive weeks	All of the above	Not reported
Gaines et al., 2018 [29]	1st and 2nd year medical students	Small group learning with librarian as the facilitator covering evidence-based medicine basics, clinical questions, searching, and matching library resources to the question.	1 or 3 weeks	All of the above	EPA 7-Ask
Getselman and White, 2011 [30]	1st year medical students.	A preassessment followed by a 30-minute lecture and a 90-minute active review of the concepts.	2 hours	All of the above	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply
Geyer and Irish, 2008 [31]	1st, 2nd, 3rd, and 4th year medical students.	A combination of web-based module consisting of tutorials and assessments, large group lecture, case-based learning, and individual assistance.	Session length varied by year.	Collaborator, Curriculum Designer, Evaluator	EPA 7-Ask, EPA 7-Acquire
Gibson and Silverberg, 2000 [32]	1st year medical students.	Seven hours of instruction over two sessions covering computer operating systems, basic computer literacy, and searching MEDLINE and the library catalog.	7 hours over 2 sessions	All of the above	EPA 7-Acquire
Gruppen et al., 2005 [33]	4th year medical students.	10 sessions of lecture and discussion on types of research literature (e.g., therapy, diagnosis, guidelines). Individual student work to generate 5 clinical questions, find evidence to answer the question, and appraise its validity	4 weeks including 10 90-minute lectures	Collaborator, Evaluator	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply

DOI: dx.doi.org/10.5195/jmla.2025.1965

		throughout the elective. Pre-post design was used to assess student knowledge and skills.			
Haley, 2019 [34]	80 total participants including 13 medical students	A single 1-hour interprofessional group book discussion facilitated by a librarian and faculty member. Pre-post survey delivered to measure interprofessional education knowledge and attitudes.	1 hour	Collaborator	EPA 7-Acquire
Hersh et al., 2002 [35]	66 total participants including 45 4th year medical students.	A large group lecture session on MEDLINE and overview of evidence-based medicine followed by two hands-on sessions applying content from the lectures 2-4 weeks after the lecture.	30 minutes large group lecture plus two 2-hour hands-on sessions.	Collaborator	EPA 7-Ask, EPA 7-Acquire
Kaplowitz and Wilkerson, 2022 [36]	1st year medical students.	A guided tour of the library, a small group review of library resources, and a large group discussion of resources.	45 minutes	Curriculum Designer	EPA 9 - Identify, EPA 9- Include, EPA 9- Establish
Kaufman et al., 1999 [37]	1st year medical students	A single introductory large-group lecture and discussion followed by four modules on evidence-based medicine resources and skills and a final project.	10 weeks	Collaborator, Curriculum Designer, Evaluator	EPA 7-Ask, EPA 7-Acquire
Lawrence and Levy, 2004 [38]	571 participants including 401 medical students.	A single workshop consisting of assessed via pre-/post-test of MEDLINE searching skills.	One session (duration not reported)	All of the above	EPA 7-Acquire
MacEachern et al., 2012 [39]	1st, 2nd, 3rd, and 4th year medical students.	A combination of lecture, case-based learning, and discussion covering various topics: Information resources (background, clinical), database searching skills.	Duration of sessions varied by year: 3 hours (1st year), 2 hours (2nd year), 2+ hours (3rd year), estimated 1 hour (4th year).	Collaborator	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply
Menard et al., 2021 [40]	3rd year medical students in their internal medicine clerkship.	Content included information resources, searching, and critical appraisal skills followed by an evidence-based medicine assignment. Intervention varied by class year but specific education strategies and methods were not reported.	14 hours instruction at the beginning of the first 2 weeks of medical school with evidence-based medicine assignments taking place during the second year.	Collaborator, Curriculum Designer, Evaluator	EPA 7-Ask, EPA 7-Acquire EPA 7 - Appraise
Minuti et al., 2018 [41]	1st and 2nd year medical students.	An interactive online tutorial covering clinical questions, searching, and information resources and a classroom session consisting of lecture and small group work.	1-2 hours	Collaborator, Curriculum Designer, Evaluator	EPA 7-Ask, EPA 7-Acquire,
Morley and Hendrix, 2012 [42]	students participating in elective course 2nd and 3rd year medical students	Combination of lecture, individual hands-on work, and group discussion.	A semester-long course of 7.5 hours	Curriculum Designer, Evaluator	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply
Muellenbach et al., 2018 [43]	1st year medical students	2 flipped classroom (pre-work, discussion, case-based learning) evidence-based medicine sessions covering an overview of EBM, information resources, clinical questions, and searching skills.	2 one-hour sessions	Collaborator, Curriculum Designer	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply
Nelson, 2018 [44]	3rd year medical students	Two online modules course covering a review of evidence-based medicine concepts, skills, and point of care resources.	Duration not reported	All of the above	EPA 7-Acquire, EPA 7-Appraise

DOI: dx.doi.org/10.5195/jmla.2025.1965

Nevius et al., 2018 [45]	US and Canadian libraries and LCME- accredited medical schools.	A twenty-eight-question survey with a mixture of qualitative and quantitative questions.	Not reported	All of the above	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise
Nicholson et al., 2019 [3]	US and Canadian health sciences libraries.	A survey assessing the extent of librarian are involved in teaching EPA 7 content, including how it is being taught, assessed, the depth of content being taught,	Not reported	All of the above	EPA 7-Ask, EPA 7-Acquire
O'Dwyer and Kerns, 2011 [46]	1st and 2nd year medical students.	Problem-based learning sessions on clinical questions, and appraising research.	12 weeks	Curriculum Designer, Evaluator	Not reported
Skhal, 2008 [47]	3rd year medical students.	Orientation session on information resources for each clinical rotation in Pediatrics, and Internal Medicine totaling 28 sessions annually followed by case-based assignment assessed by pre-post test		Leader, Collaborator, Curriculum Designer	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply
Swanberg et al., 2017 [48]	2nd year medical students	An instructional session as part of a comprehensive evidence-based medicine course covering clinical questions, searching for evidence, and appraising research.	Three-hour session consisting of a50-minute lecture followed by a mandatory 2-hour interactive lab.	Leader, Collaborator, Curriculum Designer, Evaluator	EPA 7-Ask, EPA 7-Appraise
Tagge, 2018 [49]	1st year medical students	Case-based learning covering all aspects of the evidence-based medicine process.	One 2-hour session	Leader, Collaborator, Curriculum Designer, Evaluator	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply
Thomas et al., 2020 [50]	1st and 2nd year medical students	Didactic on evidence-based medicine basics and searching PubMed followed by a small group hands-on activity.	90 minutes to 2 hours	Collaborator, Curriculum Designer, Evaluator	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply
Turner et al., 2017 [51]	1st year medical students	A single session on searching in PubMed.	A single 70-minute session	Leader, Collaborator, Curriculum Designer, Evaluator	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply
Wallach et al., 2002 [52]	1st year medical students	A mix of lecture, small group work covering finding evidence and appraising research.	Not reported	Collaborator	EPA 7-Ask, EPA 7-Acquire
Whipple et al., 2009 [53]	1st year medical students	Lecture covering background questions, using information resources to answer them, followed by case study small group exercise.	1 hour	Curriculum Designer	EPA 7-Ask, EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply
Wiecha et al., 2002 [54]	3rd year medical students	Online modules covering finding evidence, appraising research, and applying evidence to a patient.	6 weeks	Curriculum Designer, Evaluator	EPA 7-Acquire, EPA 7-Appraise
Wong and Ren, 2022 [55]	1st year medical students	A single session on library resources, advanced search strategies, and critical appraisal.	90 minutes	Leader	EPA 7-Acquire
Wrosch et al., 1998 [56]	1st year medical students	A lecture on searching in MEDLINE followed by small group work answering an assigned clinical question, and appraising an article.	A single two-hour session.	Collaborator, Curriculum Designer	EPA 7-Acquire, EPA 7-Appraise, EPA 7-Apply

DOI: dx.doi.org/10.5195/jmla.2025.1965

Zeigen and Hamilton, 2021 [57]	1st year medical students	A lecture on clinical questions and literature searching with mandatory follow-up consultation.	A single one-hour session plus mandatory consultation with a librarian.	Collaborator	EPA 7-Acquire, EPA 7-Appraise

*Domains covered are Entrustable Professional Activities (EPAs) and include stages of developing a well-formed clinical question to retrieve evidence to support clinical decision making (EPA 7: Ask, Acquire, Appraise, Apply) and collaborating as part of an interprofessional team (EPA 9: Identify, Include, Establish).

Table 2: Librarian Roles from Reviewed Studies

	Librarian Role			
	Leader	**Collaborator**	**Curriculum Designer**	**Evaluator**
Study ID				
Abate et al., 2011 [13]	X	-	X	-
Adams, 2015 [14]	X	X	X	X
Aronoff et al., 2017 [15]	-	X	X	X
Blake et al., 2018 [16]	-	X	-	-
Blanco et al., 2014 [17]	-	X	-	-
Blumenthal et al., 2005 [18]	-	X	-	-
Brahmi et al., 1999 [19]	-	X	-	-
Brown and Nelson, 2003 [20]	-	X	-	X
Burrows and Tylman, 1999 [21]	-	-	-	X
Butera et al., 2014 [22]	-	X	X	-
Cyrus et al., 2013 [23]	-	X	-	-
Dorsch et al., 2004 [24]	-	X	-	X
Earl, 1996 [25]	X	X	X	X
Eldredge et al., 1998 [26]	-	X	X	-
Eldredge et al., 2021 [27]	X	X	X	X
Gagliardi et al., 2012 [28]	X	X	X	X
Gaines et al., 2018 [29]	X	X	X	X
Getselman and White, 2011 [30]	X	X	X	X
Geyer and Irish, 2008 [31]	-	X	X	X
Gibson and Silverberg, 2000 [32]	X	X	X	X

DOI: dx.doi.org/10.5195/jmla.2025.1965

Gruppen et al., 2005 [33]	-	X	-	X
Haley, 2019 [34]	-	X	-	-
Hersh et al., 2002 [35]	-	X	-	-
Kaplowitz and Wilkerson, 2022 [36]	-	-	X	-
Kaufman et al., 1999 [37]	-	X	X	X
Lawrence and Levy, 2004 [38]	X	X	X	X
MacEachern et al., 2012 [39]	-	X	-	-
Menard et al., 2021 [40]	-	X	X	X
Minuti et al., 2018 [41]	-	X	X	X
Morley and Hendrix, 2012 [42]	-	X	-	X
Muellenbach et al., 2018 [43]	-	X	X	-
Nelson, 2018 [44]	X	X	X	X
Nevius et al., 2018 [45]	X	X	X	X
Nicholson et al., 2019 [3]	X	X	X	X
O'Dwyer and Kerns, 2011 [46]	-	-	X	X
Skhal, 2008 [47]	X	X	X	-
Swanberg et al., 2017 [48]	X	X	X	X
Tagge, 2018 [49]	X	X	X	X
Thomas et al., 2020 [50]	-	X	X	X
Turner et al., 2017 [51]	X	X	X	X
Wallach et al., 2002 [52]	-	X	-	-
Whipple et al., 2009 [53]	-	-	X	-
Wiecha et al., 2002 [54]	-	-	X	X
Wong and Ren, 2022 [55]	X	-	-	-
Wrosch et al., 1998 [56]	-	X	X	-
Zeigen and Hamilton, 2021 [57]	-	X	-	-

Outcomes and Measures

DOI: dx.doi.org/10.5195/jmla.2025.1965

Table 3 Outcome interventions from Reviewed Studies

	Outcome Interventions			
	Reaction	**Learning**	**Behavior**	**Results**
Study ID				
Abate et al., 2011 [13]	**Positive**	-	-	-
Adams, 2015 [14]	**Positive**	**Positive**	-	-
Aronoff et al., 2017 [15]	**Positive**	**Positive**	-	-
Blake et al., 2018 [16]	**Positive**	-	-	-
Blanco et al., 2014 [17]	-	-	-	-
Blumenthal et al., 2005 [18]	**Positive**	-	-	-
Brahmi et al., 1999 [19]	**Positive**	-	-	-
Brown and Nelson, 2003 [20]	-	-	**Positive**	-
Burrows and Tylman, 1999 [21]	-	**Negative**	-	**Negative**
Butera et al., 2014 [22]	-	-	-	-
Cyrus et al., 2013 [23]	-	**Positive**	-	**Positive**
Dorsch et al., 2004 [24]	**Positive**	**Positive**	**Positive**	**Positive**
Earl, 1996 [25]	-	**Positive**	-	-
Eldredge et al., 1998 [26]	-	-	-	-
Eldredge et al., 2021 [27]	-	**Positive**	**Positive**	-
Gagliardi et al., 2012 [28]	**Positive**	**Positive**	**Positive**	-
Gaines et al., 2018 [29]	-	**Positive**	**Positive**	-
Getselman and White, 2011 [30]	**Positive**	-	-	-
Geyer and Irish, 2008 [31]	**Positive**	**Positive**	**Positive**	**Positive**
Gibson and Silverberg, 2000 [32]	**Positive**	**Positive**	-	-
Gruppen et al., 2005 [33]	-	**Positive**	**Positive**	-
Haley, 2019 [34]	**Positive**	**Positive**	-	-
Hersh et al., 2002 [35]	**Positive**	**Positive**	**Positive**	-
Kaplowitz and Wilkerson, 2022 [36]	**Positive**	-	-	-
Kaufman et al., 1999 [37]	**Positive**	**Positive**	**Positive**	-

DOI: dx.doi.org/10.5195/jmla.2025.1965

Lawrence and Levy, 2004 [38]	**Positive**	**Positive**	-	-
MacEachern et al., 2012 [39]	-	-	-	-
Menard et al., 2021 [40]	**Positive**	**Positive**	**Positive**	-
Minuti et al., 2018 [41]	**Positive**	**Positive**	-	-
Morley and Hendrix, 2012 [42]	**Positive**	-	-	-
Muellenbach et al., 2018 [43]	**Positive**	-	-	-
Nelson, 2018 [44]	**Positive**	-	-	-
Nevius et al., 2018 [45]	-	-	-	-
Nicholson et al., 2019 [3]	-	-	-	-
O'Dwyer and Kerns, 2011 [46]	**Positive**	**Positive**	-	-
Skhal, 2008 [47]	**Positive**	**Positive**	-	-
Swanberg et al., 2017 [48]	**Positive**	**Positive**	-	-
Tagge, 2018 [49]	**Positive**	**Positive**	-	**Positive**
Thomas et al., 2020 [50]	**Positive**	**Positive**	-	-
Turner et al., 2017 [51]	**Positive**	-	-	-
Wallach et al., 2002 [52]	**Positive**	-	-	-
Whipple et al., 2009 [53]	**Positive**	**Positive**	-	-
Wiecha et al., 2002 [54]	**Positive**	**Positive**	-	-
Wong and Ren, 2022 [55]	-	**Positive**	-	-
Wrosch et al., 1998 [56]	-	-	-	-
Zeigen and Hamilton, 2021 [57]	-	-	-	-

Positive: findings found related to aim/research question provides positive and favorable results; Negative: findings found related to aim/research question provides negative and non-favorable results.

DISCUSSION

This scoping review found that there is a strong base of literature on the involvement of librarians in competency-based medical education. Despite this, few studies included in this review assessed outcomes related to the application of knowledge or skills taught by a librarian or used validated measures to determine the effect of the intervention. The majority of studies reported generally positive outcomes related to reaction to the intervention or knowledge retention of the content. At the same time, outcomes related to behavior of the participants or outcomes related to the application of the skills or knowledge were rarely studied.

Similar to prior reviews [4, 9-11], this scoping review found that there was a high degree of variation in how the included studies were conducted. The teaching methods, duration, setting, and assessment methods varied from study to study, making comparisons between the existing evidence challenging. This study highlights the need for more standardized interventions and assessments, especially that which could result in the understanding of the librarian's role in ensuring clinical competence among learners. When authors are writing about CBME involvement, they should include detailed descriptions

DOI: dx.doi.org/10.5195/jmla.2025.1965

about their involvement and employ more rigorous research methods to allow others to draw conclusions about efficacy.

When librarians are communicating their value to internal and external stakeholders, having landmark studies with demonstrable evidence of the contributions of librarians is essential. While librarians are publishing articles related to their involvement in competency-based medical education, existing literature may not capture the extent of work done in this area. Additional research is needed to quantify the impact of librarian involvement in competency-based medical education.

LIMITATIONS

As with any large-scale synthesis of evidence, decisions made during the design and search processes may introduce bias into the study. The decision to restrict eligibility to articles that were published or available in English and took place in LCME-accredited medical schools based in the United States potentially limited the pool of articles that could have informed our guiding questions. Additionally, hand searching of journals and conference abstracts was not conducted as part of this review due to lack of time.

DATA AVAILABILITY STATEMENT

Data associated with this article, including Excel documentation spreadsheet, are on the Open Science Framework Site for this project (https://osf.io/gcy4e).

AUTHOR CONTRIBUTIONS

John W. Cyrus: Conceptualization, methodology, investigation, visualization, writing (original draft), writing (reviewing & editing). Laura Zeigen: Conceptualization, methodology, investigation, writing (reviewing & editing). Molly Knapp: Conceptualization, methodology, investigation, (reviewing & editing). Amy E. Blevins: Conceptualization, methodology, investigation, writing (reviewing & editing). Brandon Patterson: Conceptualization, methodology, investigation, visualization, writing (reviewing & editing).

REFERENCES

1. The Core Entrustable Professional Activities (EPAs) for Entering Residency [Internet]. AAMC. [cited 2024 Aug 15]. Available from: https://www.aamc.org/about-us/mission-areas/medical-education/cbme/core-epas.
2. The Core Entrustable Professional Activities (EPAs) for Entering Residency [Internet]. AAMC. [cited 2024 Aug 15]. Available from: https://www.aamc.org/about-us/mission-areas/medical-education/cbme/core-epas
3. Nicholson J., Spak J.M., Kovar-Gough I., Lorbeer E.R., Adams N.E. Entrustable professional activity 7: opportunities to collaborate on evidence-based medicine teaching and assessment of medical students. BMC Med Educ. 2019;19(1):330.
4. Dorsch JL, Perry GJ. Evidence-based medicine at the intersection of research interests between academic health sciences librarians and medical educators: a review of the literature. Journal of the Medical Library Association : JMLA. 2012;100(4):251–7.
5. Arksey H, O'Malley L. Scoping Studies: Towards a Methodological Framework. International journal of social research methodology. 2005;8:19–32.
6. Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z. JBI Manual for Evidence Synthesis - JBI Global Wiki [Internet]. [cited 2024 Aug 15]. Available from: https://doi.org/10.46658/JBIMES-24-01
7. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MD, Horsley T, Weeks L. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Annals of internal medicine. 2018;
8. Competency-Based Medical Education (CBME) [Internet]. AAMC. [cited 2024 Feb 19]. Available from: https://www.aamc.org/about-us/mission-areas/medical-education/cbme.
9. Eldredge JD, Heskett KM, Henner T, Tan JP. Current practices in library/informatics instruction in academic libraries serving medical schools in the Western United States: a three-phase action research study. BMC medical education. 2013;13(101088679):119.
10. Maggio LA, Durieux N, Tannery NH. Librarians in Evidence-Based Medicine Curricula: A Qualitative Study of Librarian Roles, Training, and Desires for Future Development. Medical Reference Services Quarterly. 2015;34(4):428–40.
11. Swanberg SM, Dennison CC, Farrell A, Machel V, Marton C, O'Brien KK, Pannabecker V, Thuna M, Holyoke AN. Instructional methods used by health sciences librarians to teach evidence-based practice (EBP): a systematic review. Journal of the Medical Library Association. 2016;104(3):197–208.
12. Kirkpatrick DL. Evaluation of training. In: Training and development handbook. New York, NY: McGraw Hill; 1967. p. 87–112.
13. Abate LE, Gomes A, Linton A. Engaging students in active learning: use of a blog and audience response system. Medical reference services quarterly. 2011;30(1):12–8.
14. Adams NE. Asking a Great Question: A Librarian Teaches Questioning Skills to First-Year Medical Students. Medical reference services quarterly. 2015;34(4):418–27.
15. Aronoff N, Stellrecht E, Lyons AG, Zafron ML, Glogowski M, Grabowski J, Ohtake PJ. Teaching evidence-based practice principles to prepare health professions students for an interprofessional learning experience. Journal of the Medical Library Association : JMLA. 2017;105(4):376–84.
16. Blake L., Yang F.M., Brandon H., Wilson B., Page R. A Clinical Librarian Embedded in Medical Education: Patient-

DOI: dx.doi.org/10.5195/jmla.2025.1965

Centered Encounters for Preclinical Medical Students. Med Ref Serv Q. 2018;37(1):19–30.

17. Blanco MA, Capello CF, Dorsch JL, Perry G (Jerry), Zanetti ML. A survey study of evidence-based medicine training in US and Canadian medical schools. Journal of the Medical Library Association. 2014;102(3):160–8.

18. Blumenthal JL, Mays BE, Weinfeld JM, Banks MA, Shaffer J. Informatics education. Defining and assessing medical informatics competencies. Medical Reference Services Quarterly. 2005;24(2):95–102.

19. Brahmi FA, London SK, Emmett TW, Barclay AR, Kaneshiro KN. Teaching lifelong learning skills in a fourth-year medical curriculum. Medical Reference Services Quarterly. 1999;18(2):1–11.

20. Brown JF, Nelson JL. Integration of information literacy into a revised medical school curriculum. Medical reference services quarterly. 2003;22(3):63–74.

21. Burrows SC, Tylman V. Evaluating medical student searches of MEDLINE for evidence-based information: process and application of results. Bulletin of the Medical Library Association. 1999;87(4):471–6.

22. Butera G, Gomes AW, Kakar S. Expanding our roles: embedded in curriculum design. Medical reference services quarterly. 2014;33(3):292–301.

23. Cyrus JWW, Duggar DC, Woodson D, Timm DF, Mclarty JW, Pullen K, Baggett MP, Banks DE. Assessing the FACTTS: An Evidence-Based Medicine and Critical Appraisal Course for Medical Students. Medical Reference Services Quarterly. 2013;32(2):209–18.

24. Dorsch JL, Aiyer MK, Meyer LE. Impact of an evidence-based medicine curriculum on medical students' attitudes and skills. Journal of the Medical Library Association : JMLA. 2004;92(4):397–406.

25. Earl MF, Hensley K, Fisher JS, Kelley MJ, Merrick D. Faculty involvement in problem-based library orientation for first-year medical students. Bulletin of the Medical Library Association. 1996;84(3):411–6.

26. Eldredge JD, Teal JB, Ducharme JC, Harris RM, Croghan L, Perea JA. The roles of library liaisons in a problem-based learning (PBL) medical school curriculum: a case study from University of New Mexico. Health libraries review. 1998;15(3):185–94.

27. Eldredge J., Schiff M.A., Langsjoen J.O., Jerabek R.N. Question formulation skills training using a novel rubric with first-year medical students. Journal of the Medical Library Association : JMLA. 2021;109(1):68–74.

28. Gagliardi JP, Stinnett SS, Schardt C. Innovation in evidence-based medicine education and assessment: an interactive class for third- and fourth-year medical students 53. Journal of the Medical Library Association. 2012;100(4):306–9.

29. Gaines JK, Blake L, Kouame G, Davies KJ, Ballance D, Thomas Gaddy V, Gallman E, Russell M, Wood E. Partnering to Analyze Selection of Resources by Medical Students for Case-Based Small Group Learning: A Collaboration between Librarians and Medical Educators. Medical reference services quarterly. 2018;37(3):249–65.

30. Getselman A, White MS. Use of a pre-assessment tool to start a meaningful dialogue: new paradigms in library instruction. Medical reference services quarterly. 2011;30(3):245–56.

31. Geyer EM, Irish DE. Isolated to integrated: an evolving medical informatics curriculum. Medical reference services quarterly. 2008;27(4):451–61.

32. Gibson KE, Silverberg M. A two-year experience teaching computer literacy to first-year medical students using skill-based cohorts. Bulletin of the Medical Library Association. 2000;88(2):157–64.

33. Gruppen LD, Rana GK, Arndt TS. A controlled comparison study of the efficacy of training medical students in evidence-based medicine literature searching skills. Academic Medicine. 2005;80(10):940–4.

34. Haley J. Interprofessional collaboration between health sciences librarians and health professions faculty to implement a book club discussion for incoming students. Journal of the Medical Library Association. 2019;107(3):403–10.

35. Hersh WR, Crabtree MK, Hickam DH, Sacherek L, Friedman CP, Tidmarsh P, Mosbaek C, Kraemer D. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. Journal of the American Medical Informatics Association. 2002;9(3):283–93.

36. Kaplowitz J, Wilkerson L. Reaching and teaching new medical students. Academic medicine: journal of the Association of American Medical Colleges. 2002;77(11):1173.

37. Kaufman C, Conway S, Gallagher KE. Health information resources: tradition and innovation in a medical school curriculum. Medical reference services quarterly. 1999;18(1):11–23.

38. Lawrence JC, Levy LS. Comparing the self-described searching knowledge of first-year medical and dental students before and after a MEDLINE class. Medical Reference Services Quarterly. 2004;23(1):73–81.

39. MacEachern M, Townsend W, Young K, Rana G. Librarian Integration in a Four-Year Medical School Curriculum: A Timeline. Medical Reference Services Quarterly. 2012;31(1):105–14.

40. Menard L, Blevins AE, Trujillo DJ, Lazarus KH. Integrating evidence-based medicine skills into a medical school curriculum: a quantitative outcomes assessment. BMJ Evid Based Med. 2021;26(5):249–50.

41. Minuti A, Sorensen K, Schwartz R, King WS, Glassman NR, Habousha RG. Librarians Flip for Students: Teaching Searching Skills to Medical Students Using a Flipped Classroom Approach. Medical reference services quarterly. 2018;37(2):119–31.

42. Morley SK, Hendrix IC. "Information Survival Skills": a medical school elective. Journal of the Medical Library Association : JMLA. 2012;100(4):297–302.

43. Muellenbach JM, Houk KM, E Thimons D, Rodriguez B. Integrating Information Literacy and Evidence-Based Medicine Content within a New School of Medicine Curriculum: Process and Outcome. Medical reference services quarterly. 2018;37(2):198–206.

DOI: dx.doi.org/10.5195/jmla.2025.1965

44. Nelson TM. Preparing for Practice: Strengthening Third-Year Medical Students' Awareness of Point-of-Care Resources. Medical Reference Services Quarterly. 2018;37(3):312–8.

45. Nevius AM, Ettien A, Link AP, Sobel LY. Library instruction in medical education: a survey of current practices in the United States and Canada. Journal of the Medical Library Association. 2018;106(1):98–107.

46. O'Dwyer L, Kerns SC. Evolution of an information literacy curriculum for third-year medical students. Medical reference services quarterly. 2011;30(3):221–32.

47. Skhal KJ. A full revolution: offering 360 degree library services to clinical clerkship students. Medical reference services quarterly. 2008;27(3):249–59.

48. Swanberg SM, Mi M, Engwall K. An Integrated, Case-Based Approach to Teaching Medical Students How to Locate the Best Available Evidence for Clinical Care. MedEdPORTAL: the journal of teaching and learning resources. 2017;13(101714390):10531.

49. Tagge N. Leveraging accreditation to integrate sustainable information literacy instruction into the medical school curriculum. Journal of the Medical Library Association: JMLA. 2018;106(3):377–82.

50. Thomas D., Hardi A., Dufault C. Providing Information-Seeking Skills Feedback Within a Medical School Curriculum: A Partnership between Librarians and Education Specialists. Med Ref Serv Q. 2020;39(1):27–34.

51. Turner RL, M Ketchum A, Ratajeski MA, Wessel CB. Leaving the Lecture Behind: Putting PubMed Instruction into the Hands of the Students. Medical reference services quarterly. 2017;36(3):292–8.

52. Wallach PM, Roscoe L, Bowden R. The profession of medicine: an integrated approach to basic principles. Academic medicine: journal of the Association of American Medical Colleges. 2002;77(11):1168–9.

53. Whipple EC, Richwine MPW, Kaneshiro KN, Brahmi FA. Teaching first-year medical students where to go first: connecting information needs to e-resources. Medical reference services quarterly. 2009;28(2):180–6.

54. Wiecha JM, Vanderschmidt H, Schilling K. HEAL: an instructional design model applied to an online clerkship in family medicine. Academic medicine: journal of the Association of American Medical Colleges. 2002;77(9):925–6.

55. Wong H.-N., Ren L. EFFECTIVENESS OF LIBRARY INSTRUCTION ON MEDICAL STUDENTS' INFORMATION LITERACY SKILLS. BMJ Evidence-Based Medicine. 2022;27(Supplement 2):A11.

56. Wrosch JA, Morgan LK, Sullivant J, Lewis DM. Instruction of evidence-based medicine searching skills during first-year epidemiology. Medical Reference Services Quarterly. 1998;17(3):49–57.

57. Zeigen L., Hamilton A. Evolving Librarian Engagement in Undergraduate Medical Education Student Research and Scholarship. Medical reference services quarterly. 2021;40(3):337–46.

58. Ramos KD, Schafer S, Tracz SM. Validation of the Fresno test of competence in evidence based medicine. BMJ. 2003 Feb 8;326(7384):319–21.

59. Fritsche L, Greenhalgh T, Falck-Ytter Y, Neumayer HH, Kunz R. Do short courses in evidence based medicine improve knowledge and skills? Validation of Berlin questionnaire and before and after study of courses in evidence based medicine. BMJ. 2002 Dec 7;325(7376):1338–41.

60. Hirt J, Nordhausen T, Meichlinger J, Braun V, Zeller A, Meyer G. Educational interventions to improve literature searching skills in the health sciences: a scoping review. Journal of the Medical Library Association. 2020;1–5.

SUPPLEMENTAL FILES

- **Appendix A:** Search Strategies

AUTHORS' AFFILIATIONS

John W. Cyrus, cyrusjw@vcu.edu, https://orcid.org/0000-0001-5331-5151, Research and Education Librarian, Health Sciences Library, Virginia Commonwealth University, Richmond, VA

Laura Zeigen zeigenl@ohsu.edu, https://orcid.org/0000-0003-2008-0565, Health Sciences Education and Research Librarian, Oregon Health & Science University Library, Oregon Health & Science University, Portland, OR

Molly Knapp, molly.knapp@utah.edu, https://orcid.org/0000-0003-4887-8747, Training Development Manager, Network of the National Library of Medicine National Training Office, Spencer S. Eccles Health Sciences Library, Salt Lake City, UT

Amy E. Blevins, aeblevin@iu.edu, https://orcid.org/0000-0002-0859-2683, Associate Director for Public Services, Ruth Lilly Medical Library, Indiana University School of Medicine, Indianapolis, IN

Brandon Patterson, b.patterson@utah.edu, https://orcid.org/0000-0003-2432-4731, Technology Engagement Librarian, Spencer S. Eccles Health Sciences Library, University of Utah, Salt Lake City, UT

Received April 2024; accepted August 2024

ISSN 1558-9439 (Online)

ORIGINAL INVESTIGATION
DOI: dx.doi.org/10.5195/jmla.2025.1955

Making an impact: the new 2024 Medical Library Association research agenda

Marie Ascher; Margaret A. Hoogland; Karen Heskett; Heather Holmes; Jonathan D. Eldredge

See end of article for authors' affiliations.

Objective: This research project sought to identify those subject areas that leaders and researcher members of the Medical Library Association (MLA) determined to be of greatest importance for research investigation. It updates two previous studies conducted in 2008 and 2011.

Methods: The project involved a three-step Delphi process aimed at collecting the most important and researchable questions facing the health sciences librarianship profession. First, 495 MLA leaders were asked to submit questions answerable by known research methods. Submitted questions could not exceed 50 words in length. There were 130 viable, unique questions submitted by MLA leaders. Second, the authors asked 200 eligible MLA-member researchers to select the five (5) most important and answerable questions from the list of 130 questions. Third, the same 130 MLA leaders who initially submitted questions were asked to select their top five (5) most important and answerable questions from the 36 top-ranked questions identified by the researchers.

Results: The final 15 questions resulting from the three phases of the study will serve as the next priorities of the MLA Research Agenda. The authors will be facilitating the organization of teams of volunteers wishing to conduct research studies related to these identified top 15 research questions.

Conclusion: The new 2024 MLA Research Agenda will enable the health information professions to allocate scarce resources toward high-yield research studies. The Agenda could be used by journal editors and annual meeting organizers to prioritize submissions for research communications. The Agenda will provide aspiring researchers with some starting points and justification for pursuing research projects on these questions.

Keywords: Evidence Based Practice; Research; Question Formulation; Delphi Method; Research Agenda; Consensus; Leadership; Impact; Artificial Intelligence (AI)

See end of article for supplemental content.

BACKGROUND

The Medical Library Association (MLA) research policy directs the MLA Research Caucus (formerly the Research Section) to identify "research priorities in the field" due to the limited research capacity of its membership. [1-2] MLA Research Caucus's Research Agenda Committee has implemented a research protocol for identifying these research priorities. The Research Agenda Committee in 2008 conducted a Delphi study that produced the top-ranked research questions at that time [3]. The Research Agenda Committee conducted a second Delphi study in 2011 that included refinements to the protocol [4] and published a supplemental inventory of all submitted questions [5]. The 2011 Delphi study led to a Systematic Review Project involving teams that pursued systematic or scoping reviews aimed at answering many of the top-ranked questions [6-7]. Thus, the previous MLA research agendas enabled researchers in our profession to focus their investigations on answering those specific answerable research questions that MLA leaders and researchers had identified as most important. The present study sought to improve upon past research agenda protocols while updating the research priorities in a new MLA Research Agenda.

METHODS

Phase One

Phase One of this Delphi method project began on September 7, 2023, with a Qualtrics survey of 495 MLA elected and appointed leaders to identify those research questions considered to be most important for the profession.

DOI: dx.doi.org/10.5195/jmla.2025.1955

MLA leaders on the national level were defined as all elected officials, all chairs and members appointed to national level committees, and the co-editors of *JMLA*. At the caucus and chapter level, leaders were defined as all elected officers and appointed committee chairs. The authors included MLA caucus and chapter leaders to ensure the diversity of the racial, ethnic, library type, professional function, geographic location, age, life experience, and perspectives represented by these leaders. Names and email addresses were obtained from rosters on MLA Board, committee, caucus, and chapter pages, or the online MLA Member Directory. **Table 1** lists the categories of leaders with the numbers of officials filling these categories. Some of the 495 leaders served in multiple roles so the authors had to de-duplicate the leaders' names. Only those leaders whose names appeared in the MLA Member Directory were eligible to participate in Phase One of the study.

Each leader was instructed to submit one question. The wording of the email to these leaders read: "Thank you for agreeing to participate in this brief survey on 'What is the most important answerable research question facing the profession?' Please enter your single sentence, most important answerable question in the following text box. Your question must be on a single topic and not exceed fifty (50) words." Following three reminder emails sent to the same 495 MLA leaders, Phase One of the study ended on September 27, 2023.

Table 1 MLA Leaders included in Phase One

Role	Number
National Officers	5
National Board of Directors	12
National Editors	2
National Committee Chairs	46
National Committee Members	128
Caucus Officers	235
Caucus Committee Chairs	21
Chapter Officers	92
Chapter Committee Chairs	104
Subtotal	**645**
Duplicates removed	147
RAC members removed	3
Final Total	**495**

During October 2023, one author [KH] organized the 130 viable and de-identified questions, phrases, or single-word topics submitted by the MLA leaders into broad themes. **Table 2** lists the subject themes with numbers of questions in each category with another column recording those questions also related to Artificial Intelligence (AI). The other four authors reviewed these categorized questions and discussed the subject categories among themselves and the appropriateness of specific questions in these categories. The authors reasoned that the grouping of questions into broad themes would reduce respondents' survey fatigue, making Phase Two voting easier by grouping similar questions from the long list. This is the sole manipulation of the questions and done specifically to make the voting process easier and minimize bias based on question order. Questions were reproduced in the Phase Two survey exactly how they were submitted in Phase One. Only those MLA leaders who both submitted questions and listed their names with email addresses in Phase One became eligible to participate again in Phase Three, as described later in this Methods section.

Table 2 Themes in Phase One Submitted Questions

Question Theme	Number of Questions	AI-related Questions
Future Casting	15	9
Information literacy/Data literacy/AI Literacy/Misinformation	12	3
Measuring the impact of librarian work	8	0
Retention/recruitment/professional development	27	2
Scholarly communication and Collections	9	0
Value of or role of librarian *(broadly defined)*	17	1
Value of or role of librarian, using measurement, supported by data, or strategies for indicating value/role	17	0
Value of or role of librarian within technological changes	25	18

Phase Two

Phase Two involved surveying MLA-member researchers to identify what they think are both the most important and researchable questions of those submitted in Phase One. The authors defined researchers in this study as MLA member colleagues who published peer reviewed research articles on health sciences librarianship during the years 2019 through August 2023. Additionally, as a change in protocol from previous iterations of this process,

DOI: dx.doi.org/10.5195/jmla.2025.1955

aimed at more inclusivity, MLA Research Training Institute graduates 2019-2022 and MLA research award recipients 2019-2023 were included as well. **Table 3** lists the composition of the Phase Two participants.

MLA member published researchers included in this study published in selected core journals. The authors defined these core journals as having peer reviewed research articles on topics that would be in-scope for MLA members:

Evidence Based Library and Information Practice; EBLIP

Health Information and Libraries Journal

Hypothesis: Research Journal for Health Information Professionals

Journal of Electronic Resources for Medical Libraries; JERML

Journal of Hospital Librarianship

Journal of the Medical Library Association; JMLA

Medical Reference Services Quarterly.

One investigator [MH] searched the LISTA (Library, Information Science and Technology Abstracts) database (EBSCO) for the publication years of 2019 to the present in the core journals on September 6, 2023. All articles from these six journals during this time period were reviewed for eligibility. They were found using a simple journal title search of the source field. Since only articles specifically related to health sciences librarianship were reviewed from *EBLIP,* a subject search was created for this journal title. The search strategy for these relevant articles combined searching for the journal title in the source field with the following search string: health science OR health science libraries OR health information professionals OR informationist OR informationists OR health science librarian OR solo librarian OR hospital librarian OR hospital OR health system OR health center OR health centre OR embedded librarian. In total, all of the searches combined produced 1,010 article entries that were uploaded to the Rayyan™ screening platform for review.

Table 3 MLA-Member Researchers included in Phase Two

Role	Number
MLA Research Paper Authors	123
MLA Research Award Winners	28
RTI Participants, not published	69
RTI Published Authors	20
Subtotal	**240**
3 RAC members removed	3
Duplicates Removed	37
Deduplicated Total	**200**

Two authors [MH, JE] identified from those results the *research* articles published in the six core journals 2019 to August 2023. These two authors defined research as the "critical and exhaustive investigation or experimentation, having for its aim the discovery of new facts and their correct interpretation, the revision of accepted conclusions, theories, or laws in the light of newly discovered facts, or the practical application of such new or revised conclusions, theories, or laws" [8]. To be included the articles had to have an identifiable research method with measurable results for their authors to be included in this study. The authors designated over 100 "maybe" entries in their Rayyan screening platform that required a direct examination of the item to determine whether or not it fit the definition of research. The easiest items to exclude were editors' introductions, resource reviews, errata, article appendices, letters, commentaries, editorials, course descriptions, narrative reviews, or background articles. Expert consensus statements were excluded unless they contained a substantive research component. Case reports had to have a methods section, some data, and ideally a "lessons learned" section. The authors excluded surveys with fewer than 50 respondents and no measurable results. They also excluded any methods articles with fewer than 1,200 words and 10 references. History, biography, or obituary articles had to have be least 1,000 words in length and have at least 5 references.

Once the research articles in the six core journals were identified, all author names were extracted. MLA member authors were then identified using the MLA Member Directory and email addresses were recorded by two authors of the present study [HH, MA].

MLA members who received MLA Research Awards for the years 2019 through 2023 were pooled with the published researchers. Finally, those MLA member colleagues who completed the MLA Research Training Institute 2019-2023 also were added to this pool of researchers. MLA members who were identified as Leaders in Phase One were not eliminated from Phase Two. The total 200 unique researchers in this pool were invited on November 3, 2023 to participate in this second phase of the Delphi process by voting for five (5) of the 130 Phase One questions on the basis of both the (1) "importance of these questions" and (2) the "feasibility of answering these research questions." Following three emailed reminders for the identified researchers to cast their votes, the Phase Two survey closed on February 1, 2024.

Phase Three

In Phase Three, the 130 MLA leader participants who had submitted questions in Phase One had the final vote in determining the questions for inclusion in Research Agenda. Those participating leaders were asked by email on February 8, 2024, to vote on their top five (5) questions

DOI: dx.doi.org/10.5195/jmla.2025.1955

from the top-ranked 36 questions produced from the researchers' votes in Phase Two. Each time a potential respondent opened the survey they encountered a new randomly ordered list of 36 questions to diminish either primacy, [9] recency bias, [10] or response order bias [11-12] brought on by the sequence of questions. Phase Three ended on March 4, 2024, less than 10 months since the MLA Research Caucus Executive Board and the MLA Board of Directors had approved the study protocol.

RESULTS

This three-phase Delphi method study produced the 15 top-ranked research questions comprising the new 2024 MLA Research Agenda that appear in **Table 4**. Phase One of the study generated 130 questions from MLA leaders. Phase Two resulted in 36 questions selected by MLA-member researchers from those 130 questions as the most answerable given available research methods. Phase Three asked the participating leaders from Phase One to select their top five choices among the 36 questions that emerged from the MLA-member researchers in Phase Two. Questions 14 and 15 on Table 4 are so similar that they might be merged into one research question. The questions after Question 15 on the full list seemed repetitive. The full list of the 130 originally submitted and de-identified questions can be found in the supplemental files accompanying this article. The questions in **Table 4** have been edited to fix only punctuation and capitalization errors present in the submitted questions.

Table 4 Final 15 Questions of the new MLA Research Agenda - Results of Phase Three

	Question	# of Votes
1	What is the most effective way to demonstrate the impact of librarians on health sciences research, education, and patient care?	39
2	In heavily data-driven academic medical centers and hospitals, what data should be collected and how should it be displayed and analyzed to continue to justify our value to stakeholders, including CEOs and CFOs?	30
3	What is the knowledge gap between new graduates from accredited library schools and the skills needed to work in medical libraries?	29
4	Do clinical medical librarians, by serving on rounds, provide a measurable impact on patient care (length of stay reduction, readmission reduction, etc.)?	26
5	How can we engage with diverse populations to pursue careers in health sciences librarianship?	26
6	Because so many of the people we serve don't understand what we can do or how much we can help them, how can we more effectively and actively demonstrate our value to them in a persuasive way?	24
7	Does librarian integration into health sciences instruction positively impact information seeking behaviors of health sciences trainees and professionals?	22
8	Do health sciences libraries and librarians have any measurable (statistically significant) positive impacts on consumer health, the outcomes of medical care, the productivity of biomedical researchers and the knowledge obtained by graduates of biomedical and health sciences training programs, and at what total cost?	22
9	What medical library services are most important now and what will be most important in the near future as information technology continues to rapidly evolve?	22
10	How do services provided by medical librarians contribute to the achievement of a larger institution's goals?	22

DOI: dx.doi.org/10.5195/jmla.2025.1955

11	How will we address the fundamental changes to scholarly publishing and library budgets that are occurring with the rise of Open Science?	20
12	How can we restructure our professional organizations to meet the networking and continuing education needs of the average early career librarian via regional chapter collaborations versus a national meeting that is financially out of reach for most early and mid-career professionals?	18
13	How do we best measure long-term learning outcomes related to library-taught competencies (e.g. EPA 7) in health sciences curricula?	17
14	How will generative AI impact the health sciences librarianship profession?	17
15	How will current and future developments in artificial intelligence affect our profession - both negatively or positively?	17

DISCUSSION

Many of the final 15 questions are about impact. One class of questions seeks to gauge the impact of external influences upon our profession, such as the emergence of artificial intelligence, which appears twice in the final 15 questions (Q14, Q15) and new technologies (Q9 and new publishing paradigms (Q11). Another class of question asks whether health information professionals can have a measurable impact upon outcomes at their broader institutions, such as in research (Q1, Q8), education (Q1, Q7, Q8, Q13), and patient care (Q1, Q4, Q8), as well as the persistent need to demonstrate value to health care leaders (Q2, Q6, Q10). The classic Rochester Study [13] and the Detroit Study [14-15] were early attempts to answer these kinds of impact questions. Methodologically, measuring impact in broadly-defined studies makes it difficult to account for potential confounders. Lastly, two of the final 15 questions relate to the education and recruitment of new health information professionals (Q3, Q5) including the need to attract more diverse body of librarians to health sciences librarianship (Q5) and the role of professional organization in networking and continuing education (Q12). Many of the questions are similar, with some updated nuances, to those asked in 2012. All de-identified data sets for this study can be accessed in the Supplemental Files.

The next steps of the project will be to promote the new 2024 MLA Research Agenda through established MLA communication channels and to engage with colleagues who wish to join teams organized to answer one each of the 15 top-ranked research questions. The authors will facilitate the formation of these teams but will not explicitly coordinate these efforts. These new teams might wish to conduct original research, systematic reviews, or scoping reviews to address their chosen research question. These initial teams might even break into smaller teams to narrow their focus. The questions on **Table 4** inevitably will need to be refined and, in most cases, narrowed or broken down into multiple more discrete questions to be suitable for research. The new research agenda questions could be used by annual meeting organizers to recruit paper or poster presentation topics. The leading journals in our field might invite prospective researcher authors to submit manuscripts on selected top-ranked research questions. Editors might rate manuscripts on whether the research study addressed one of the top-ranked questions. The new 2024 MLA Research Agenda provides aspiring researchers with starting points and the rationale for implementing their research.

LIMITATIONS

There are a few limitations related to this study. For one, the process was conducted during a specific period of history in the US with concerns in the larger society that might have had an outsized effect on leaders' submitted questions and the subsequent phases of votes. As only one example, after months of anxiety about possible career displacements [16-18] following OpenAI's Fall 2022 release of Chat GPT amplified in news media, it should come as no surprise that 33 artificial intelligence-related questions appeared on the initial 130 submissions. This historic artifact [19] was reduced slightly by the two subsequent rounds of voting much later in 2023 and early 2024.

Additionally, during the early portion of Phase Two an attempt to use Javascript to randomize the order of questions (in order to reduce related bias) led to a corruption of some voter output. Researchers who submitted during this phase were invited to resubmit responses. It is unknown how many of those early voters resubmitted their votes.

Lastly, some questions that remained in the study were too broadly or too vaguely stated to serve as productive questions for researchers to pursue with any known research study designs. Phase Two's inclusion of researchers culled some of these questions from further consideration. A number of the final 15 research questions

DOI: dx.doi.org/10.5195/jmla.2025.1955

will need further refinement and reframing by teams conducting original research, systematic reviews, or scoping reviews based upon a specific question. Leaders who submitted questions in Phase One were instructed to limit their single-sentence questions to no more than fifty (50) words which was less than the previous sixty (60) in the prior iteration. Future iterations of this kind of Delphi study should include additional guidelines and more detailed guidance for formulating truly answerable research questions for participants in Phase One. The investigators might want to recommend even shorter word limits to submitted questions.

CONCLUSION

This Delphi study has produced a broad consensus statement on what subjects should be elevated in priority in the next five years. The Agenda will provide aspiring researchers with some starting points and justification for pursuing research projects on these questions. The 15 research questions in **Table 4** potentially will guide leadership and researcher collective efforts in multiple contexts to build the evidence base needed by our professional colleagues.

IRB STATEMENT

The University of New Mexico Health Sciences Human Protection Program approved the study 23-241 on July 12, 2023.

ACKNOWLEDGMENTS

The authors would like to acknowledge the hundreds of MLA members who contributed questions and voted in the course of this study. We wish you success in your pursuit of answers to these questions.

DATA AVAILABILITY STATEMENT

The data associated with this project will be made available as supplemental files: the research protocol, the initial 130 questions in full, and the de-identified spreadsheets including voting tallies.

AUTHOR CONTRIBUTIONS

Marie T. Ascher: Project administration, methodology, survey management, data collection and curation, and production of tables and substantive contributions to the text during review and editing.

Margaret A. Hoogland: Identified the published authors of research articles, data curation, writing: review & editing.

Karen M. Heskett: Revisions to study methodology, identifying and organizing themes from questions submitted during Phase One, data collection of MLA members and emails during Phase One, and writing: review and editing.

Heather N. Holmes: Launching the new agenda project, refining the research methodology, verifying MLA membership, writing, and editing the manuscript.

Jonathan D. Eldredge: Conceptualized and operationalized the original Delphi study design and then later revised it in conjunction with the other authors, secured IRB approval, communicated with study participants, identified the published authors of research articles, analyzed the study results, and wrote the first draft of the manuscript.

REFERENCES

1. Medical Library Association. The MLA research imperative executive summary [Internet]. The Association; 2007. [cited 10 Sep 2024]. https://www.mlanet.org/wp-content/uploads/2024/07/2007_mla_research_imperative.pdf.
2. Grefsheim SF, Rankin JA, Perry GJ, McKibbon KA. Affirming our commitment to research: the Medical Library Association's research policy statement: the process and findings. J Med Libr Assoc. 2008 Apr;96(2):114–20. DOI: http://dx.doi.org/10.3163/1536-5050.96.2.114
3. Eldredge JD, Harris MR, Ascher MT. Defining the Medical Library Association research agenda: methodology and final results from a consensus process. J Med Libr Assoc 2009;97(3):178-185. DOI:10.3163/1536-5050.97.3.006.
4. Eldredge JD, Ascher MT, Holmes HN, Harris MR. The new Medical Library Association research agenda: final results from a three-phase Delphi study. J Med Libr Assoc 2012;100(3):214-218. DOI:10.3163/1536-5050.100.3.012.
5. Harris MR, Holmes HN, Ascher MT, Eldredge JD. Inventory of research questions identified by the 2011 MLA research agenda Delphi study. Hypothesis 2013 Winter; 24(2): 5-16. [cited 10 Sep 2024]. https://journals.indianapolis.iu.edu/index.php/hypothesis/issue/view/1624/610.
6. Ascher MT, Eldredge JD. MLA Research Section's Research Agenda Committee Systematic Review Project: a status report. Hypothesis 2017 Spring/Summer; 29 (1): 5-17. DOI: https://doi.org/10.18060/25208.
7. Boden C, Ascher MT, Eldredge JD. Learning while doing: program evaluation of the Medical Library Association systematic review project. J Med Libr Assoc 2018;106(3):284-293. doi:10.5195/jmla.2018.286.
8. Gove PB, Merriam-Webster, Inc. Webster's third new international dictionary of the English language, unabridged. Springfield, MA.: Merriam-Webster; 1993.
9. Baron J. Thinking and deciding. Cambridge University Press, 1988: 259-61.
10. Plous S. The psychology of judgment and decision making. McGraw-Hill, 1993.

DOI: dx.doi.org/10.5195/jmla.2025.1955

11. Moore DW. Measuring new types of question order effects: additive and subtractive. Public Opinion Quarterly 2002; 66: 80-91. DOI: https://doi.org/10.1086/338631.

12. Carp FM. Position effects on interview responses. Journal of Gerontology 1974; 29 (5): 581-87. DOI: https://doi.org/10.1093/geronj/29.5.581.

13. Marshall JG. The impact of the hospital library on clinical decision making: the Rochester study. Bull Med Libr Assoc 1992;80(2):169-178.

14. Klein MS, Ross FV, Adams DL, Gilbert CM. Effect of online literature searching on length of stay and patient care costs. Acad Med 1994;69(6):489-495. DOI:10.1097/00001888-199406000-00017.

15. Klein MS, Ross F. End-user searching: impetus for an expanding information management and technology role for the hospital librarian. Bull Med Libr Assoc 1997;85(3):260-268.

16. Gertner J. Wikipedia's moment of truth. N Y Times Mag. 2023 July 18. [cited 10 Sep 2024]. https://www.nytimes.com/2023/07/18/magazine/wikipedia-ai-chatgpt.html.

17. Metz C, Schmidt G. Elon Musk and others urge a pause in A.I., citing 'profound risks to society.' New York Times 2023 March 29. [cited 10 Sep 2024]. https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html.

18. Huang K. Alarmed by AI chatbots, universities start revamping how they teach. New York Times 2023 Jan 16. [cited 10 Sep 2024]. https://www.nytimes.com/2023/01/16/technology/chatgpt-artificial-intelligence-universities.html

19. Cook TD, Campbell DT. Quasi experimentation: design & analysis issues for field settings. Houghton Mifflin Company; 1979. p.51-52, 74.

SUPPLEMENTAL FILES

- **Appendix A**: Phase 1 Survey Responses Report
- **Appendix B**: Phase 2 Survey MLA Research Agenda
- **Appendix C**: Phase 3 Final Tallies MLA Research Agenda 2024
- **Appendix D**: Phase 3 MLA Research Agenda Delphi Study 2024

AUTHORS' AFFILIATIONS

Marie Ascher, MS, MPH, marie.ascher@nymc.edu, https://orcid.org/0000-0001-6394-1041, Lillian Hetrick Huber Endowed Director, Phillip Capozzi, MD Library, and Assistant Professor, School of Medicine and Department of Epidemiology, New York Medical College, New York, NY

Margaret A. Hoogland, MLS, AHIP, Margaret.Hoogland@UToledo.edu, https://orcid.org/0000-0002-9932-3605, Associate Professor and Clinical Medical Librarian, University Libraries, The University of Toledo, Toledo, OH

Karen Heskett, MST, kheskett@ucsd.edu, https://orcid.org/0000-0003-4501-6172, Medical Librarian, University of California–San Diego Library, San Diego, CA

Heather N. Holmes, MLIS, AHIP, holmesh@musc.edu, https://orcid.org/0000-0002-2318-440X, Associate Director of Libraries and Professor, Academic Affairs Faculty, The Medical University of South Carolina, SC

Jonathan D. Eldredge, PhD, AHIP, FMLA,jeldredge@salud.unm.edu, https://orcid.org/0000-0003-3132-9450, Professor, Health Sciences Library and Informatics Center and Family & Community Medicine, School of Medicine; Associate Program Director, Clinical Informatics, Graduate Medical Education; University of New Mexico, Albuquerque, New Mexico

Received March 2024; accepted October 2024

ISSN 1558-9439 (Online)

ORIGINAL INVESTIGATION
DOI: dx.doi.org/10.5195/jmla.2025.1860

Automated tools for systematic review screening methods: an application of machine learning for sexual orientation and gender identity measurement in health research

Ashleigh J. Rich; Emma L. McGorray; Carrie Baldwin-SoRelle; Michelle Cawley; Karen Grigg; Lauren B. Beach; Gregory Phillips II; Tonia Poteat

See end of article for authors' affiliations.

Objective: Sexual and gender minority (SGM) populations experience health disparities compared to heterosexual and cisgender populations. The development of accurate, comprehensive sexual orientation and gender identity (SOGI) measures is fundamental to quantify and address SGM disparities, which first requires identifying SOGI-related research. As part of a larger project reviewing and synthesizing how SOGI has been assessed within the health literature, we provide an example of the application of automated tools for systematic reviews to the area of SOGI measurement.

Methods: In collaboration with research librarians, a three-phase approach was used to prioritize screening for a set of 11,441 SOGI measurement studies published since 2012. In Phase 1, search results were stratified into two groups (title with vs. without measurement-related terms); titles with measurement-related terms were manually screened. In Phase 2, supervised clustering using DoCTER software was used to sort the remaining studies based on relevance. In Phase 3, supervised machine learning using DoCTER was used to further identify which studies deemed low relevance in Phase 2 should be prioritized for manual screening.

Results: 1,607 studies were identified in Phase 1. Across Phases 2 and 3, the research team excluded 5,056 of the remaining 9,834 studies using DoCTER. In manual review, the percentage of relevant studies in results screened manually was low, ranging from 0.1 to 7.8 percent.

Conclusions: Automated tools used in collaboration with research librarians have the potential to save hundreds of hours of human labor in large-scale systematic reviews of SGM health research.

Keywords: Sexual and Gender Minorities; Health; Methods; Systematic Review; Automation

See end of article for supplemental content.

INTRODUCTION

Sexual and gender minority (SGM) populations disproportionately experience poor health compared to heterosexual and cisgender populations. For example, SGM populations experience increased risk for physical and mental health issues such as depression, anxiety, HIV, and some cancers [1, 2], with research suggesting that these disparities are related to experiences of minority stress (e.g., stigmatization, discrimination, negative internalized attitudes) in relation to one's SGM identity [3, 4]. While existing research makes clear that these disparities exist, understanding the extent and nature of these disparities requires comprehensive, accurate measurement of sexual orientation and gender identity (SOGI). Accurate and consistent measurement of SOGI helps researchers to paint the clearest picture of the health inequities faced by SGM populations. Advancing this understanding is necessary to develop interventions to promote SGM health equity.

Existing SOGI measurement strategies often fall short of providing the information needed to fully understand SGM disparities. One issue is a lack of standardized validated measurement across health research and practice contexts, which prevents straightforward integration of findings from different settings. Existing measurement approaches often do not capture the

DOI: dx.doi.org/10.5195/jmla.2025.1860

multidimensionality of sexual orientation, a construct that includes attraction, behavior, and identity [5]. Sex and gender are often conflated, captured in a limited capacity via one step item (i.e. 'male', 'female', 'transgender') instead of best practice two-step measures (i.e. a sex assigned at birth item plus a current gender identity item) [6]. Further, gender and sex are often treated as binary constructs encompassing only identities such as "man" and "woman" or "male" and "female," reinforcing notions of gender and sex that prevent nonbinary and intersex identities from being appropriately measured [7]. The lack of pre-existing sampling frames as well as the historical exclusion of SGM people from routine public health surveillance and other health research efforts constitute other challenges [8].

Even ongoing efforts to address these inconsistencies and offer recommendations for standardized SOGI measurement can replicate limitations of prevailing measurement strategies. Importantly, the recently released US National Academies of Science, Engineering, and Medicine (NASEM) landmark 2022 report, *Measuring Sex, Gender Identity and Sexual Orientation* [9], systematically evaluating SSOGI measurement in the US, providing measurement guidance, and setting related research priorities for the NIH and beyond, is limited by gender identity measurement recommendations that may conflate sex and gender and erase non-binary identities as well as fail to capture sexual orientation multidimensionality. To better understand these issues and get a comprehensive view of measurement of SOGI in health research, we undertook a systematic review. Conducting a systematic review in SGM health poses a number of challenges. First, opportunities for SGM health research are growing [10], producing a large body of research results to screen when conducting systematic reviews. Second, searching for research related to SOGI measurement involves key terms likely to be found in a wide range of studies, including studies completely unrelated to SGM health or SOGI measurement. This means that searching for research in this area is likely to produce a large amount of research irrelevant to researchers' questions, increasing the time needed to screen search results.

One potential solution to this problem is the use of automated tools such as machine learning, which have long been used to minimize the time and labor needed to screen the large volume of search results that arises when investigating complicated or wide-ranging research questions [11]. However, despite these tools' potential, [12–18], they have not often been leveraged to streamline the process of conducting systematic reviews [11].

Unfamiliarity with machine learning and other automated tools may be one barrier to implementation of these tools in systematic reviews. However, librarians have access to the training, expertise, and software needed to conduct effective searches and screen results using automated tools [11]. Collaborations with librarians trained in automation tools pose a promising opportunity for research teams to effectively use these tools to ensure high-quality, efficient reviews, and we established such a collaboration in the current research. As part of a larger project reviewing and synthesizing how SOGI has been assessed within the health literature, we provide an example of the application of automated tools for systematic reviews to the area of SOGI measurement.

METHODS

Team Roles

The University of North Carolina Health Sciences Library (UNC HSL) offers both consulting and co-authoring services to affiliated researchers. As co-authors, librarians lead the construction of search strategies, perform the searches, advise on automation tools, maintain an EndNote Library, set up the review within Covidence, and contribute to the manuscript. The non-librarian researchers co-design and review the search strategy, screen the studies in both the title/abstract and full text stages, assess quality of included studies, synthesize research, and write the review.

Search Methods

The search strategy, developed by the research team and librarians, included controlled vocabulary terms and keywords based on the concepts of a) sexual and gender minorities (e.g., gay, lesbian, bisexual, transgender) and b) measurement (**Table 1**). Health sciences librarians conducted comprehensive searches in four bibliographic databases: PubMed (NLM), CINAHL (EBSCOhost), PsycInfo (EBSCOhost), and Health and Psychosocial Instruments-HAPI (EBSCOhost). Based on the volume of the results, availability of potential databases, and the indexing of the known journals of interest, the team selected subject-specific databases that would be most likely to contain relevant results. The search was limited to English-language documents with a published date of 2012 or later. Since the field of SGM health research has exploded in the past decade, SGM literature reviews with longer timeframes ultimately include research since 2010 [19], and SOGI measurement prior to the recent past likely includes discredited findings, the team applied a date filter to focus on the state of the SOGI literature in the past decade. The search included peer-reviewed journal articles reporting primary data focused on SSOGI measurement in health research, conducted in the United States. Conference abstracts, case reports, editorials, reviews, and any other non-peer reviewed literature were excluded from eligibility.

DOI: dx.doi.org/10.5195/jmla.2025.1860

Table 1 Measurement-related terms searched in title used to stratify search results.

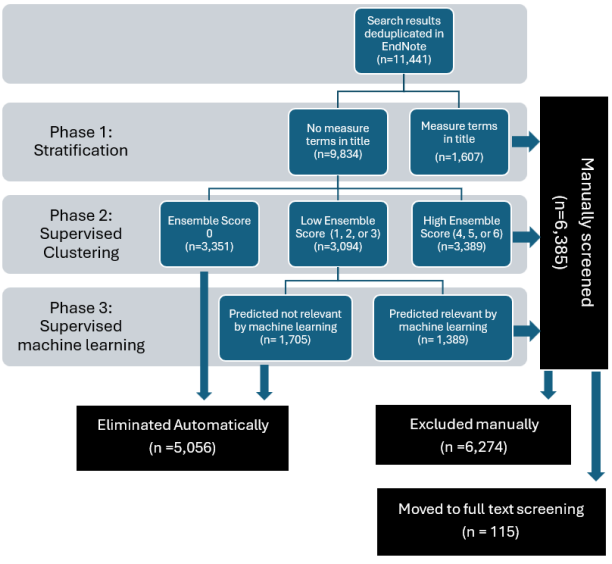
Root term searched in title field	Terms captured
Instrum*	Instrument; Instrumental; Instruments; Instrumentation
Measur*	Measure; Measures; Measured; Measurement; Measurements; Measuring
Scal*	Scale; Scales; Scaled; Scaling
Surv*	Survey; Surveyed; Surveys; Surveying; Surveil; Surveillance; Surveilling
Valid*	Valid; Validate; Validates; Validated; Validating; Validation; Validity

*indicates truncation to capture alternate word forms.

Prioritization of Literature for Manual Screening

Studies most likely to be relevant from the search results were prioritized for manual screening in three phases described below (**Figure 1)**. Citations were then manually screened for inclusion at the title and abstract level, then at the full text level, by two independent subject matter experts using Covidence Systematic Review Software [20].

Figure 1 Summary of screening methodology by phase



Phase 1: Stratification

Following de-duplication in EndNote, the search results were stratified into two groups. Studies with one or more measurement-related terms in the title (**Table 1**) were identified in EndNote and screened manually in Phase 1 and the remaining studies moved forward to Phase 2. The team chose to stratify the results because we expected that studies with measurement-related terms in the title were more likely to make measurement a focus of the paper, rather than an incidental mention in the abstract. Cawley noted that using a stratified approach can be helpful to ensure a subset of results are all considered in manual review before the application of automation tools such as machine learning [11].

Phase 2: Supervised Clustering Using DoCTER

In Phases 2 and 3, results not screened in Phase 1 were prioritized with Document Classification and Topic Extraction Resource (DoCTER) [21]. DoCTER uses publicly available clustering and machine learning algorithms to prioritize search results using the text of titles and abstracts, including K-means, non-negative matrix factorization (NMF), Naïve Bayes, linear support vector machines (linear SVC), and k-nearest neighbor (KNN). Varghese et al. provides details on these conventional machine learning algorithms as used by DoCTER [22].

In Phase 2, supervised clustering—a form of semi-supervised learning that groups an unclassified corpus of studies and a set of known relevant (i.e., "seed") studies into clusters based on text similarities in titles and abstracts—was used. Seed studies are a form of training data but require fewer positive studies than typically necessary for machine learning algorithms. Ideally, a target of 25-50 seeds should be identified by reviewing a random subset of search results.

Clusters containing seed studies are likely to contain relevant unclassified studies. Clusters are prioritized for manual screening based on the number of seed studies they contain until a desired recall target is reached. For example, if 100 seed studies are used and 95% recall is desired, then clusters are prioritized for manual review until 95 or more of the seeds are captured.

Seeds (positive training data) should be identified at random from the unclassified corpus to avoid selection bias and to produce accurate predictions of recall. Ideally, subject matter experts should screen studies at random to select at least 25 seeds. Negative training data are not necessary for supervised clustering. Varghese, Cawley, and Hong provide further details on supervised clustering and demonstrate that the method rivals accuracy rates of supervised machine learning algorithms while requiring less training data [22]. Cawley provides summary data for a series of case studies using the approaches outlined here

DOI: dx.doi.org/10.5195/jmla.2025.1860

by librarians at UNC HSL including stratification and prioritizing studies for screening in a two-phased approach with supervised clustering and supervised machine learning [11].

The ensemble approach to supervised clustering uses two algorithms: k-means and nonnegative matrix factorization (NMF) and three cluster sizes: 10, 20, and 30. Using each algorithm with the three different cluster numbers yields six different clustering models (e.g., KM-10 model is the k-means algorithm with 10 clusters and KM-20 is the k-means algorithm with 20 clusters). The six models were applied to title and abstract text of the citations not screened in Phase 1, along with a set of seed studies.

The output of supervised clustering with a six-model ensemble approach is an ensemble score (ES) for each study that ranges from 6 to 0. The ES indicates the number of models where the study was found in a cluster prioritized by DoCTER. Citations with ES = 6 are predicted to have a higher likelihood of relevance compared to studies with lower ensemble scores. Citations with an ES = 0 are not predicted relevant by any of the six models and are typically excluded without manual screening.

Phase 3: Supervised Machine Learning Using DoCTER

In Phase 3, results less likely to be relevant (ES = 3, 2, or 1) from Phase 2 were further prioritized using supervised machine learning. The decision to move to supervised machine learning is recommended when precision (i.e., the number of relevant studies as a percentage of all studies screened manually) starts to diminish rapidly. Moving to machine learning to prioritize studies further allows for more studies to be excluded without manual screening.

Supervised machine learning uses different algorithms than clustering (e.g., naïve Bayes, support vector machines) and requires a relatively large training dataset. Whereas supervised clustering requires approximately 25-50 relevant studies for training data, machine learning requires positive and negative training data. The amount of training data needed varies based on many factors but from experience we endeavor to use at least 100 positive studies. The sizes of training datasets used for this approach range from the low hundreds (van de Bulk et al.) to high thousands (Liao et al.). Cawley et al. ran three simulations of a similar application of machine learning and used approximately 200 positive studies for training data in each of the three simulations and reached 95% recall in each instance [23, 24].

After running the supervised machine learning process in DoCTER, each study is given a probability score based on how likely it is to be relevant. Unlike supervised clustering with an ensemble approach, which puts studies into batches, machine learning algorithms provide a probability score for each individual study. The training data for supervised machine learning were derived from studies manually screened in Phases 1 and 2. Cawley provides evidence that a two-step approach of supervised clustering followed by supervised machine learning is effective at reducing the manual screening burden without significantly impacting recall of relevant articles and that training data for supervised machine learning can be drawn from labelled data in earlier steps [11].

RESULTS

Search Results

In total, 17,814 citations were returned from all databases searched. Results were imported to EndNote and duplicates were removed. After removing duplicates, 11,441 citations were prioritized for manual screening.

Phase 1 Results

Phase 1 included all results with measurement-related terms in title (**Table 1**), identified by a keyword search in EndNote. All 1,607 results in this group were screened manually, given that these studies had a higher likelihood of being relevant (**Figure 1**) and 85 relevant studies were identified during this step.

Phase 2 Results

Studies not containing a measurement-related term in title (n = 9,834) were moved to Phase 2 and prioritized with DoCTER [21] software using supervised clustering with an ensemble approach (**Figure 1**).

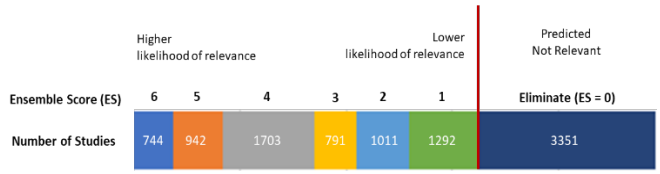
Prior to Phase 1, the research team screened the titles and abstracts of 500 studies, selected at random from the search results, to identify seeds. As noted above, seeds should be identified from a random sample of the unclassified corpus to avoid selection bias and allow for accurate predictions of recall. In this step, 39 studies were classified as relevant by subject matter experts and used as seeds to prioritize the 9,834 results not screened in Phase 1.

In Phase 2, supervised clustering with an ensemble approach was used to prioritize results for manual screening. In total, 6,483 results had an ES = 1 or higher and were retained for either manual screening or further prioritization. A total of 3,351 results had an ES = 0 (**Figure 2**) and were excluded without manual screening. For Phase 2 of screening, studies with an ES = 4 or higher were screened manually (n = 3,389) (**Figure 1**). Only 10 relevant studies were found in these results. This very low precision is unusual but was not unexpected by the research team. The nature of the systematic review question necessitated a broad search strategy that would result in a large number of false positives. Due to the very low precision for studies with ES = 6, 5, or 4, the research

DOI: dx.doi.org/10.5195/jmla.2025.1860

team further prioritized the remaining studies with an ES = 1, 2, or 3 in Phase 3 using supervised machine learning.

Figure 2 Supervised clustering results



Note: Studies with ensemble score (ES) of greater than 1 were considered for manual screening or further prioritization.

Phase 3 Results

For Phase 3 of screening, supervised machine learning was applied to further prioritize studies with an ES = 3, 2, or 1 (n = 3,094). Training data for supervised machine learning were derived from screening results of Phases 1 and 2. The machine learning algorithm prioritized a total of 1,389 studies as likely to be relevant using a recall threshold of 95%. These studies were manually screened, and all remaining studies (n = 1,705) were excluded without manual review (Phase 3; **Figure 1**). Studies were screened in order of probability score in descending order from most likely to be relevant to least likely to be relevant.

Of the 1,389 studies screened manually in Phase 3, only 11 relevant studies were identified after full-text screening was completed (1 study was excluded after full text screening). The 11 studies identified in Phase 3 were found in the top 25% of 1,389 studies screened for Phase 3 when ordered by probability of being relevant. The bottom 75% of studies (n = 1,063) did not contain any relevant studies. This provides evidence that the approach was effective and that few, if any, additional relevant studies were likely to be found in the studies excluded without manual screening.

DISCUSSION

In this study, a machine learning approach was applied to literature screening in the conduct of a systematic review of SOGI measurement research. This work provides a practical application of automated methods to systematic reviews in the context of SOGI measurement and SGM health, illustrating that automated tools can help researchers to efficiently use time and labor resources. Such considerations are especially important in fast-growing areas such as SGM health and SOGI measurement where low-precision searches will likely remain normative; this study serves as a potential model for researchers in these areas. Nearly all health domains have fast growing areas of research (e.g., emerging infectious diseases) or topics where the historical volume of literature consistently poses a challenge any time a new research question is asked in the domain (e.g., tobacco- and HIV-related research).

This project also illustrates the utility of collaborations between research teams and health science librarians when conducting systematic reviews, as librarians have training in the required skillset and access to the necessary software to implement automated tools [11], enabling research partners to focus on their disciplinary and content area expertise.

The application of machine learning to systematic literature reviews is most often for the literature screening step [25]. In this study, the research team screened a total of 6,385 studies manually. Using supervised clustering and supervised machine learning in Phases 2 and 3 allowed us to exclude 5,056 studies without manual screening.

At all phases of manual review, search precision was very low and ranged from 0.1 to 7.8 percent with the highest precision in Phase 1. Overall search precision was 1.8%, which was consistent with the research team's expectations of relatively high sensitivity and low specificity given the growing SGM health research literature and relatively sparse research in SOGI measurement. The risk of misclassification is low as SOGI terms (i.e. sexual orientation, gender identity) are very specific to SGM research and not used in other disciplines. Given the low search precision following manual screening for all three phases, studies with an ES = 0 were excluded from manual screening. Tran et al. note that reducing the number of citations that must be screened manually using automation may not be recommended for reviews assessing efficacy of clinical interventions but may be acceptable in other instances [26]. Further, it is notable that using automation to reduce the number of citations that must be screened manually may allow research teams to develop broader research questions and contribute to a paradigm shift in how relevant literature is found [24, 26].

Using machine learning to exclude studies without manual review carries the risk of Type 2 errors (i.e., false negatives). Saving time and resources is the tradeoff to missing relevant studies. Consensus is that a recall threshold of 95% is an acceptable level of risk for systematic reviews using AI-assisted screening methodology [27, 28]. DoCTER and other similar applications allow the user to specify the recall threshold which is estimated using training data. Given the statistical underpinnings of the stopping criteria, we are confident we missed 5% or fewer of the relevant studies [11].

When available, simulation data bears this out and we consistently find 95% or higher recall using this

DOI: dx.doi.org/10.5195/jmla.2025.1860

methodology on simulated data [11, 24]. With simulated data we use a fully labelled dataset and simulate the performance of these approaches to confirm that we can reach the desired recall threshold of 95%. The authors also recommend building safeguards into the process to reduce the number of Type 2 errors when possible, including supplementing the keyword search with handsearching, soliciting expert knowledge, and reviewing bibliographies of relevant preprints or recent articles.

One major strength of this study was the efficiency the automated approach afforded, and which other researchers can hopefully achieve by adopting similar approaches. Researchers have estimated that screening a title and abstract takes about two minutes of human labor across two screeners [29], meaning that excluding over 5,000 studies from manual review alone saved over 160 hours of researcher time. In the event screening is completed by paid research assistants, this may correspond to important budget impacts, a key consideration in the responsible stewardship of research funds. For example, the savings would be a minimum of $2,500 based on the standard hourly wage of $18-20 for Research Assistants at Northwestern University, where the study was conducted. The incorporation of human expertise was essential in the use of automated methods in this study; specifically, human experts guided feature selection, model development, and result validation and stratified some items for manual screening to optimize the use of automated tools. Overall, the timesaving achieved from application of the automated approach to screening was especially useful given that the search was low in precision, a challenge that other SGM researchers are also likely to encounter when conducting systematic reviews. Researchers in other areas or with narrower-scope research topics may achieve higher-precision results when using these methods.

Although the search was low precision even after prioritizing studies with machine learning, this was not unexpected given the nature of our constructed search strategy, as SGM health research has been expanding [30], searches of related topics have been similarly high-volume [31], and little attention has been paid to SOGI measurement relative to the total body of research on SGM health. The low-precision search does not undermine the utility of the automated approach, as without this approach, screening results would have been more resource-intensive. However, there is still significant room for improvement in precision when using automation to identify relevant literature. Large language models (LLMs) such as those incorporated into generative AI tools from Google, OpenAI, and Anthropic show potential for improving precision in the application of article screening [26].

In future work, researchers should consider applying machine learning tools to test these approaches in other areas of SGM health, COVID-19, HIV and other infectious diseases, and tobacco research to aid in identifying other contexts in which use of these methods might be most useful. Using a machine learning approach for future systematic reviews—and incorporating partnerships with experienced librarians when doing so—has the potential to ensure that researchers can efficiently search, review and synthesize the literature to make the most comprehensive and well-informed recommendations for future research and practice.

FUNDING

This work was supported by the National Institute on Alcohol Abuse and Alcoholism under grant number R01AA029076 (PIs: Beach, Phillips). AJR is supported by a Canadian Institutes of Health Research Fellowship (Grant #429552).

CONFLICT OF INTEREST

We have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The database search strategies are available in searchRxiv https://www.cabidigitallibrary.org/doi/10.1079/searchRxiv.2024.00566.

AUTHOR CONTRIBUTIONS

AJR contributed to project conceptualization, analysis, investigation, methodology, project administration and supervision, as well as leading draft writing, review and editing. ELM contributed to draft writing, review and editing. CBS, MC, and KG contributed to methodology, analysis, investigation, visualization, and writing. LB, GP, and TP contributed to conceptualization, funding acquisition, administration, and reviewed and edited manuscript drafts.

REFERENCES

1. The Health of lesbian, gay, bisexual, and transgender people: building a foundation for better understanding. Choice Reviews Online. 2012 Jan 1;49(05):49-2699-49–2699.
2. Mayer KH, Bradford JB, Makadon HJ, Stall R, Goldhammer H, Landers S. Sexual and Gender Minority Health: What We Know and What Needs to Be Done. Am J Public Health. 2008 Jun;98(6):989–95.
3. Brooks V. Minority stress and lesbian women. Lexington Books; 1981.
4. Meyer IH. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: Conceptual issues and research evidence. Psychological Bulletin. 2003 Sep;129(5):674–97.

DOI: dx.doi.org/10.5195/jmla.2025.1860

5. Rich AJ, Armstrong HL, Cui Z, Sereda P, Lachowsky NJ, Moore DM, Hogg RS, Roth EA. Sexual orientation measurement, bisexuality, and mental health in a sample of men who have sex with men in vancouver, canada. Journal of Bisexuality. 2018 Jul 3;18(3):299–317.
6. Restar AJ, Lett E, Menezes NP, Molino AR, Poteat TC, Dean L, Glick JL, Baker KE, Cole SW. Getting Precise about Gender and Sex Measurement: A Primer for Epidemiologists. Am J Epidemiol. 2024 Jun 29;kwae144.
7. Hyde JS, Bigler RS, Joel D, Tate CC, van Anders SM. The future of sex and gender in psychology: Five challenges to the gender binary. American Psychologist. 2019 Feb;74(2):171–93.
8. Rich AJ, Lachowsky NJ, Sereda P, Cui Z, Wong J, Wong S, Jollimore J, Raymond HF, Hottes TS, Roth EA, Hogg RS, Moore DM. Estimating the size of the MSM population in Metro Vancouver, Canada, using multiple methods and eiverse data sources. Journal of urban health. 2018 Apr;95(2):188–95.
9. Committee on Measuring Sex, Gender Identity, and Sexual Orientation, National Academies of Sciences, Engineering, and Medicine (NASEM). Measuring sex, gender identity, and sexual orientation [Internet]. Washington, D.C.: National Academies Press; 2022 [cited 2022 Mar 14] p. 26424. Available from: https://www.nap.edu/catalog/26424.
10. Notice of Special Interest (NOSI): Research on the Health of Sexual and Gender Minority (SGM) Populations - NOT-MD-22-012 - NIDDK [Internet]. National Institute of Diabetes and Digestive and Kidney Diseases. [cited 2023 Aug 15]. Available from: https://www.niddk.nih.gov/research-funding/current-opportunities/not-md-22-012.
11. Cawley MA. Using Machine Learning to Locate Evidence More Efficiently: New Roles for Academic Research Librarians. In: Handbook of Research on Academic Libraries as Partners in Data Science Ecosystems [Internet]. IGI Global; 2022 [cited 2022 Dec 19]. p. 144–68. Available from: https://www.igi-global.com/chapter/using-machine-learning-to-locate-evidence-more-efficiently/www.igi-global.com/chapter/using-machine-learning-to-locate-evidence-more-efficiently/302752.
12. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. 2015 Dec;4(1):5.
13. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019 Dec;8(1):163, s13643-019-1074–9.
14. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, Marshall IJ. 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. 2021 May 1;133:140–51.
15. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics. 2010 Dec;11(1):55.
16. Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, Kelly MP, Thomas J. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Research Synthesis Methods. 2014;5(1):31–49.
17. Mostafa J, Lam W. Automatic classification using supervised learning in a medical document filtering application. Information Processing & Management. 2000 May 1;36(3):415–44.
18. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. Journal of the American Medical Informatics Association. 2006 Mar 1;13(2):206–19.
19. Rich AJ, Scheim AI, Koehoorn M, Poteat T. Non-HIV chronic disease burden among transgender populations globally: A systematic review and narrative synthesis. Prev Med Reports. 2020 Dec;20:101259.
20. Veritas Health Innovation. Covidence Systematic Review Software [Internet]. Melbourne, Australia: Veritas Health Innovation; 2022. Available from: www.covidence.org
21. Document Classification and Topic Extraction Resource [Internet]. Virginia, USA: ICF; 2017. Available from: https://www.icf-docter.com/.
22. Varghese A, Cawley M, Hong T. Supervised clustering for automated document classification and prioritization: a case study using toxicological abstracts. Environ Syst Decis. 2018 Sep 1;38(3):398–414.
23. van den Bulk LM, Bouzembrak Y, Gavai A, Liu N, van den Heuvel LJ, Marvin HJP. Automatic classification of literature in systematic reviews on food safety using machine learning. Current Research in Food Science. 2022 Jan 1;5:84–95.
24. Cawley M, Vest T, Carlson R, Eckel S. Machine-assisted literature screening for a medication use process related systematic review. American Journal of Health-System Pharmacy. 2024 Jul;[In press].
25. Bolanos F, Salatino A, Osborne F, Motta E. Artificial Intelligence for Literature Reviews: Opportunities and Challenges [Internet]. arXiv; 2024 [cited 2024 Jul 1]. Available from: http://arxiv.org/abs/2402.08565
26. Tran VT, Gartlehner G, Yaacoub S, Boutron I, Schwingshackl L, Stadelmaier J, Sommer I, Alebouyeh F, Afach S, Meerpohl J, Ravaud P. Sensitivity and Specificity of Using GPT-3.5 Turbo Models for Title and Abstract Screening in Systematic Reviews and Meta-analyses. Ann Intern Med. 2024 Jun;177(6):791–9.
27. Callaghan MW, Müller-Hansen F. Statistical stopping criteria for automated screening in systematic reviews. Syst Rev. 2020 Dec;9(1):273.
28. Liao J, Ananiadou S, Currie LG, Howard BE, Rice A, Sena SE, Thomas J, Varghese A, Macleod MR. Automation of citation screening in pre-clinical systematic reviews [Internet]. 2018 [cited 2024 Jun 27]. Available from: http://biorxiv.org/lookup/doi/10.1101/280131
29. Cawley M, Beardslee R, Beverly B, Hotchkiss A, Kirrane E, Sams R, Varghese A, Wignall J, Cowden J. Novel text analytics approach to identify relevant literature for human health risk assessments: A pilot study with health effects of

DOI: dx.doi.org/10.5195/jmla.2025.1860

in utero exposures. Environment International. 2020 Jan;134:105228.

30. Jackman KB, Bosse JD, Eliason MJ, Hughes TL. Sexual and gender minority health research in nursing. Nursing Outlook. 2019 Jan;67(1):21–38.

31. Marshall Z, Welch V, Minichiello A, Swab M, Brunger F, Kaposy C. Documenting Research with Transgender, Nonbinary, and Other Gender Diverse (Trans) Individuals and Communities: Introducing the Global Trans Research Evidence Map. Transgender Health. 2019 May;4(1):68–80.

SUPPLEMENTAL FILES

- **Appendix A**: Recognize Methods

AUTHORS' AFFILIATIONS

Ashleigh J. Rich, Ashleigh.rich@duke.edu, School of Nursing, Duke University, Durham, NC

Emma L. McGorray, EmmaMcgorray2023@u.northwestern.edu, Department of Psychology, Northwestern University, Evanston, IL

Carrie Baldwin-SoRelle, chbs@email.unc.edu, Health Sciences Library, University of North Carolina Chapel Hill, Chapel Hill, NC

Michelle Cawley, mcawley@email.unc.edu, Health Sciences Library, University of North Carolina Chapel Hill, Chapel Hill, NC

Karen Grigg, kgrigg@email.unc.edu, Health Sciences Library, University of North Carolina Chapel Hill, Chapel Hill, NC

Lauren B. Beach, lauren.beach@northwestern.edu, Department of Medical Social Sciences, Department of Preventative Medicine, Northwestern University, Evanston, IL

Gregory Phillips II, Glp2@northwestern.edu, Department of Medical Social Sciences, Department of Preventative Medicine, Northwestern University, Evanston, IL

Tonia Poteat, tonia.poteat@duke.edu, School of Nursing, Duke University, Durham, NC

Received October 2023; accepted October 2024

ISSN 1558-9439 (Online)

ORIGINAL INVESTIGATION
DOI: dx.doi.org/10.5195/jmla.2025.1936

Algorithmic indexing in MEDLINE frequently overlooks important concepts and may compromise literature search results

Alexandre Amar-Zifkin; Taline Ekmekjian; Virginie Paquet; Tara Landry

See end of article for authors' affiliations.

Objective: To evaluate the appropriateness of indexing of algorithmically-indexed MEDLINE records.

Methods: We assessed the conceptual appropriateness of Medical Subject Headings (MeSH) used to index a sample of MEDLINE records from February and March 2023. Indexing was performed by the Medical Text Indexer-Auto (MTIA) algorithm. The primary outcome measure is the number of records for which the MTIA algorithm assigned subject headings that represented the main concepts of the publication.

Results: Fifty-three percent of screened records had indexing that represented the main concepts discussed in the article; 47% had inadequacies in the indexing which could impact their retrieval. Three main issues with algorithmically-indexed records were identified: 1) inappropriate MeSH assigned due to acronyms, evocative language, exclusions of populations, or related records; 2) concepts represented by more general MeSH while a more precise MeSH is available; and 3) a significant concept not represented in the indexing at all. We also noted records with inappropriate combinations of headings and subheadings, even when the headings and subheadings on their own were appropriate.

Conclusions: The indexing performed by the February-March 2023 calibration of the MTIA algorithm, as well as older calibrations, frequently applied irrelevant or imprecise terms to publications while neglecting to apply relevant terms. As a consequence, relevant publications may be omitted from search results and irrelevant ones may be retrieved. Evaluations and revisions of indexing algorithms should strive to ensure that relevant, accurate and precise MeSH terms are applied to MEDLINE records.

Keywords: Abstracting and Indexing, Algorithms, Medical Subject Headings, PubMed, MEDLINE, Search Strategies, Database Searches, Information Storage and Retrieval, MeSH

INTRODUCTION

The purpose of indexing, or "the act of describing or identifying a document in terms of its subject content" [1], is to make pertinent documents retrievable. Controlled vocabulary terms provide indexers with specific, preferred entries for concepts that can manifest in multiple, synonymous ways, and have been deployed in many different bibliographic databases and research domains [2]. The Medical Subject Headings (MeSH) thesaurus, developed by the United States National Library of Medicine (NLM) in 1960, is the controlled vocabulary of terms "used for indexing, cataloging, and searching for biomedical and health-related information and documents" [3], and is used in the NLM catalog and in MEDLINE, the premier bibliographic database in biomedical sciences.

For decades, countless health and information professionals have taught and been taught that using MeSH in their literature searches in MEDLINE will increase the precision of their queries, thus improving the relevance of their results [4, 5]. A classic demonstration of the usefulness of MeSH is that a study excluding patients with diabetes would not receive a MeSH term indicating diabetes, whereas a text word search for 'diabetes' would irrelevantly retrieve that publication if it had mentioned its exclusion criteria in the title or abstract. Searchers are taught that MeSH terms assigned to a record reflect the main concepts of an article and thus reliably indicate its most important aspects [6-8]. As evidence of the presumed value and utility of controlled vocabulary such as MeSH to indicate aboutness, resources like the Cochrane Handbook, which also influence searching practices more broadly, direct searchers to use them [9].

Indexing of biomedical literature has been moving gradually from manual or human-based indexing towards automatic semantic indexing for more than a decade [10]. The movement to automation aims to reduce time-to-indexing and cost, identified in Mao and Lu [11] as 2-3

DOI: dx.doi.org/10.5195/jmla.2025.1936

months and $10 per publication, respectively, and must be considered in the context of the ever-increasing volume of biomedical literature. For example, 1,279,327 indexed citations were added to MEDLINE in 2023, compared to 734,052 in 2013, a 74% increase over 10 years [12].

In December 2021, the NLM announced its intention that "all citations indexed for Medline will be indexed by MTIA" (Medical Text Indexer-Auto) [13]. Although algorithmic indexing has provided indexing suggestions to human indexers since 2002 via the Medical Text Indexer (MTI) [14], the move towards fully automated indexing significantly changes how bibliographic records are indexed [15]. Humans with subject matter expertise previously determined central elements of articles and selected appropriate MeSH terms for them; as of April 2022, the front line of indexing for all records is performed by an algorithm, with humans limiting their curation to sets involving genes and proteins [13].

Briefly, the MTIA algorithm determined which MeSH terms to apply to a record by:

- identifying uncommon or specific terms in that article's title and abstract;
- finding MeSH for those terms;
- gathering MeSH which have been assigned to other records with similar uncommon or specific terms from within MEDLINE; and
- ranking these terms and deciding which to apply to the record [16].
    - The MTIA used several processes to rank terms. One such step involved the prioritization of a subheading over a MeSH heading when a term was present in both thesauri (e.g.: the subheading *pharmacokinetics* would be preferred over the MeSH term *Pharmacokinetics*). (*See Section 13. MH/SH Substitution*) [16].

Unlike human indexers, MTIA did not consider the journal in which an article appeared, the author-suggested keywords, or the full text of the article, any of which might provide more insight into concepts germane to the aboutness of the article. For example, studies have found that the methods section can also be quite informative, often containing information on species, sex, and age groups, all of which are required by MeSH indexing guidelines as check tags [17, 18].

Previous research has identified issues with indexing in MEDLINE. Two such studies, Minguet et al [19] and Tonin et al [20], found that indexing in pharmacy publications did not consistently use MeSH terms from the Pharmacy branches of the MeSH tree, while Portaluppi [21] arrived at a similar conclusion for articles on chronobiology. Layton and Clarke [22] found that the representation of statistical concepts in interventional dentistry publications was lacking, and Wilczynski and Haynes [23] identified issued with the consistency and accuracy of indexing of knowledge syntheses. Moore, Yaqub and Sampat [24] explored the classification of documents by disease area and Neveol et al [8] and Rae et al [25] identified challenges in pairing main headings and subheadings.

A more recent study sought to assess the outputs of algorithmic indexing by taking a sample of records published in the year 2000, before human indexers began to receive indexing recommendations from the MTI, inputting those records into a public-facing MTI, and comparing the MTI indexing to that of humans [26]. While this identified some issues with MTI outputs, notably in headings representing populations ('check tags') and the influence of sentence structure on concept ranking, it is ultimately a comparison of algorithmic indexing output to human indexing, and only lightly questions the appropriateness of the terms assigned by the algorithm. By contrast, our study seeks to determine whether automatically assigned index terms reflect the main concepts of an article and indicate its most important aspects.

In the months preceding this research, the authors each encountered multiple MEDLINE records where the indexing (later determined to be automated) did not align with their experience or expectations. For example, for the article, *An exploratory study on support for caregivers of people with vision impairment in the UK* [27] there is no indexing representing visually impaired persons, or even of visual impairment. By contrast, *Laparoscopic versus open elective right hemicolectomy with curative intent for colon adenocarcinoma* [28] is indexed with the MeSH term *Child, Preschool* for no apparent reason, with no other MeSH indicating the correct age range of study participants.

The first rule of indexing is to include all topics known to be of interest to users that are treated substantively within the document [29]. Although some indexing theorists argue that there can be no single "correct" set of index terms for a document, as different requesters may seek out the same document for different reasons [30], we would argue that, given the purpose of indexing – to make documents retrievable – it is reasonable to assert that the *essential* topics, or concepts, of a document should be accurately represented in its indexing.

These most recent steps towards fully automated and algorithmic indexing, even with human spot-checks, raise fundamental questions:

- How well do algorithmically-applied MeSH indicate a publication's essential concepts?
- When information specialists and health professionals run MeSH-only searches, expecting that MeSH will identify relevant research on their

DOI: dx.doi.org/10.5195/jmla.2025.1936

topics, is the algorithm causing anything to be overlooked? If so, what?

If algorithmically applied MeSH are found to frequently overlook or misrepresent salient concepts of publications, how much confidence should instructors of literature searching to tomorrow's health professionals have in their teaching of MeSH as a first step to quickly and precisely identifying the best available evidence?

We recognize that as of late April 2024, MTIA has been replaced by MTIX [Medical Text Indexer-NeXt Generation], which uses a machine-learning algorithm rather than a rules-based algorithm [31]. Nonetheless, as searchers and researchers continue to engage with recent, and therefore algorithmically-indexed, publications, we hope that the following analyses and insights into automated indexing lay a foundation for a deeper and broader understanding of indexing algorithms and their outputs.

METHODS

This study assessed whether a sample of MTIA-indexed articles were indexed with MeSH that adequately described their main concepts. As such, this research should not be interpreted as comparing algorithmically-indexed records to human-indexed records.

We piloted our screening process with a set of 10 records drawn from January 2022 and 2023. Our team of librarians, each with significant expertise in database searching and instruction, independently assessed each record and determined whether the assigned MeSH terms adequately represented main concepts. We found that our team was generally able to identify similar main concepts for each article within the set, and likewise determine whether the assigned MeSH was concordant with those concepts. However, the team struggled to identify major concepts in articles that were not based in clinical health sciences, stemming instead from the broader constellations of life sciences such as genetics and zoology, as well as civil engineering, agricultural sciences and software development. We therefore decided to allow screeners to exclude these articles when screening our final set.

Our final sample was sourced using Ovid MEDLINE® ALL. On March 31, 2023, we sampled 20 days of MEDLINE between February 6 and March 7, 2023, skipping weekends and holidays so that each date had a similar number of publications. We opted for a recent date range because the algorithm can be recalibrated over time, and we wanted to assess recently indexed publications rather than records that might have been indexed using a previous configuration.

We only used records with "automated" in the Indexing Method [IG] field. This excluded records with no value in the IG field, indicating entirely human indexing, and records with a value of "curated", indicating revision by human indexers of terms applied by MTIA. We used a random sequence generator [32] to select 50 records from each date, and used Ovid's internal deduplication function to remove duplicate records. Our final sample consisted of 998 unique records. Sample queries for date and randomization are presented below:

1. 20230206.ed and automated.ig and medline.st and english.la (3420)
2. from 1 keep [50 unique random numbers between 1-3420] (50)

We exported the 998 records into a spreadsheet. Each record was assigned to two screeners, who were blinded to each other's work and to the MeSH terms that had been assigned to each record by the MTIA. Because our final sample was 100x larger than our test set, we established a two-step screening protocol. The first pass of screening allowed for a record to be excluded if there was insufficient information in the record—for example, if there was no abstract and the title was uninformative – or if the screener considered the article to be outside the scope of clinical health sciences.

An example of a record both without an abstract and with an uninformative title is the article *Blue as an Orange* [33]. An example of a record that was considered to be outside the scope of clinical health sciences is *Mathematical analysis of topological and random m-order spread models* [34].

In the second pass of screening, each screener assigned concepts based on the available data in the record, either using MeSH headings of which they were already aware (e.g.: Patient Education As Topic) or more general terminology (e.g.: air quality or air pollution). Screeners were not limited in the number of concepts they could assign, so long as they felt the concepts were descriptive of an aspect of the publication. As Publication Types are not conceptual, per se, we did not direct screeners to apply publication types like 'Randomized Controlled Trial', but we did encourage the use of 'As Topic' terms when appropriate (e.g. 'Randomized Controlled Trial As Topic'). Screeners were instructed to consider the article they were screening as a 'seed' or 'target' article, and assign concepts that they felt would be part of a search for the article, as well as articles on a similar topic, within the database.

Once a screener had identified the main concepts of a record, they were un-blinded to the MeSH assigned to the article. The screener would then indicate agreement or disagreement between their concepts and the MeSH assigned by MTIA. In cases of uncertainty or ambiguity, MeSH scope notes were consulted as needed to ensure understanding of a term was accurate. Three abridged rows from the screening tool, showing assessments from one screener, are provided in Table 1. The "yes" in the Agreement? column indicates that the concepts identified by the screener were deemed present in the indexing. Our

DOI: dx.doi.org/10.5195/jmla.2025.1936

complete dataset is available online at https://osf.io/ckj3m/.

Disagreements between screeners–where one screener indicated that their concepts and the MeSH aligned, while those of the other screener did not–were resolved through discussion by two authors (AA-Z and TL), who were blinded to the identities of the disagreeing screeners. The results of the analysis are presented below.

RESULTS

From our sample of 1,000 records, 2 duplicates were removed.

From the 998 records which were screened, we excluded 287 because they did not contain enough information to assign concepts or because they were outside what we considered as the scope of health sciences. The remaining 711 records were screened by our team. The flow of record screening is presented in the PRISMA-styled figure below.

After resolving disagreements, we found that 377 records (53%) had been assigned MeSH terms that adequately represented the main concepts present in the title and abstract and 334 records (47%) had one or more deficiencies in their indexing. The team found that these 334 records had commonalities which we have grouped into four main categories, with varying potential impacts on retrieval. These are summarized in Table 2.

Table 1 Sample records in screening form

Record	Screener-identified concepts	MeSH assigned by MTIA	Agreement?
Indications for continuous electroencephalographic (cEEG) monitoring: What do they tell us? *Epilepsy Research* [35]	EEG Epilepsy Length of monitoring?	Female Humans Middle Aged Male Prospective Studies Epilepsy/di [Diagnosis] *Epilepsy Monitoring, Physiologic Electroencephalography *Lupus Erythematosus, Systemic	Yes
Development and validation of a nomogram for evaluating the incident risk of carotid atherosclerosis in patients with type 2 diabetes. *Frontiers in Endocrinology* [36]	Carotid atherosclerotic disease Type 2 diabetes Risk	Humans *Diabetes Mellitus, Type 2 Nomograms *Non-alcoholic Fatty Liver Disease *Carotid Artery Diseases Risk Factors	Yes
The trends and determinants of seasonal influenza vaccination after cardiovascular events in Canada: a repeated, pan-Canadian, cross-sectional study. *Health Promotion and Chronic Disease Prevention in Canada* [37]	Canada flu vaccination / vaccination cardiovascular diseases public policy	Humans Canada/ep [Epidemiology] Cross-Sectional Studies Influenza, Human/ep [Epidemiology] Influenza, Human/pc [Prevention & Control] *Influenza, Human Seasons Vaccination	No (missing cardiovascular diseases)

DOI: dx.doi.org/10.5195/jmla.2025.1936

Figure 1

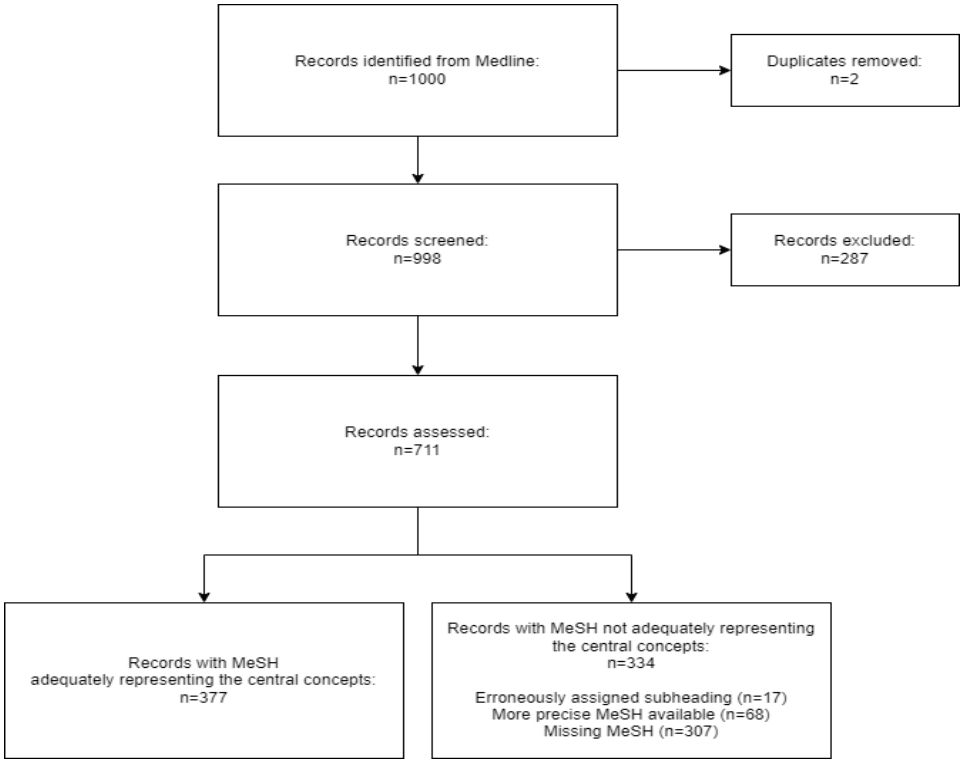


Table 2 Results

Indexing issues	Number of records*	% of records assessed	Example
Appropriate subheadings erroneously assigned to a main heading	17	2.4%	Safety and tolerability of obeticholic acid in chronic liver disease: a pooled analysis of 1878 individuals [38]. The subheading of drug therapy is attached to the MeSH for pruritis, the adverse effect of obeticholic acid discussed in the abstract. This linkage implies that the subject of the article is drug therapy for the adverse effect, rather than drug therapy for a condition (Non-alcoholic Fatty Liver Disease/dt [Drug Therapy]), an adverse effect of the drug (Chenodeoxycholic Acid/ae [Adverse Effects]), and a chemically-induced adverse effect (Pruritis/ci [Chemically Induced])
Concept represented by MeSH while a more precise MeSH was available	68	9.6%	Ambulatory oxygen therapy in lung transplantation candidates with idiopathic pulmonary fibrosis referred for pulmonary rehabilitation [39] Indexed with Oxygen, without the therapeutic use subheading and without the more precise Oxygen Inhalation Therapy heading.
Significant concept not represented in the indexing at all	307	43.2%	Bridging knowledge gaps in paediatric chronic urticaria through a video-based educational tool [40]

DOI: dx.doi.org/10.5195/jmla.2025.1936

			There is no representation of anything relating to education in the indexing, despite Patient Education as Topic being available.
MeSH terms adequately represented the main concepts present in the title and abstract	377	53.0%	Development and Validation of the HIV-CARDIO-PREDICT Score to Estimate the Risk of Cardiovascular Events in HIV-Infected Patients [41] HIV infections, Cardiovascular Diseases, and Risk were all represented in the indexing.

*While 334 records had indexing issues, some had more than one.

DISCUSSION

This study identified a number of indexing issues for publications that were indexed using MTIA. Three of the issues we identify may result in work arounds that could significantly and undesirably increase the number of retrieved citations. For instance, researchers conducting knowledge syntheses could bypass the issue of appropriate subheadings being erroneously assigned to a main heading by using 'floating' subheadings in their searches (to retrieve any record indexed with a specific subheading, regardless of which heading it is attached to).

For the 9.6% of records identified, where records were represented by broader MeSH while more precise MeSH terms were available, searchers would need to include these broader terms in their search queries in order to retrieve these publications. For example, when searching for articles about the experiences of visually impaired persons, rather than simply searching for Visually Impaired Persons, a searcher now needs to include less precise terms from the Vision Disorders and Eye Diseases MeSH trees, potentially going further to Eye, Optometry, Ophthalmology, Optometrists or Ophthalmologists. Although health educators, information professionals [6] and the NLM itself [42] have explicitly or tacitly recommended searching by the most precise MeSH, this may result in the exclusion of articles which have been indexed using MTIA. In practical terms, this means that, because the algorithm cannot "read between the lines" searchers must replace a precise MeSH heading with MeSH terms tangentially related to their topic.

For records in which a significant concept was not represented in the indexing, the impact on searching is clear. Because they were not assigned appropriate headings for a central concept, a MeSH only search, no matter how expansive, will not retrieve those records because they were not assigned appropriate headings for a central concept. Guo, Gotz and Wang [43] conceptualize the omission of a relevant concept from an article's indexing as a bottleneck in the search. As a consequence, a searcher would only find relevant publications through title-abstract-keyword searching, which may introduce noise into search results.

In the course of our screening, our team founds records where the MTIA assigned somewhat florid and unusual indexing terms that did not pertain to the concepts present in the title or abstract. As these findings were incidental and would likely not result in the exclusion of relevant publications from a search, we did not systematically note them. Nevertheless, we felt it was important to discuss them here, as other health information professionals may have encountered similar instances in their own search results.

For example, we found several instances of MTIA erroneously assigning a MeSH term based on the use of an acronym or evocative language in the title or abstract. For instance, the article *Bridging knowledge gaps in paediatric chronic urticaria through a video-based educational tool* [40] was indexed with the MeSH term Copper, likely because the authors abbreviated 'chronic urticaria' as CU (the chemical symbol for copper). Likewise, we found that the use of metaphor, simile or rhetoric to describe or illustrate ideas sometimes led to indexing errors, such as the article *Not all cauliflowers are HPV: challenge* [44] being assigned the MeSH term Brassica, despite being about cauliflower-shaped genital warts and not the noble brassica family.

We also found several instances of irrelevant indexing as a result of the MTIA assigning MeSH terms based on the indexing of similar records within the database ("neighboring records") [16]. This type of error is illustrated by the indexing of *Prevalence of undernutrition and its associated factors among older adults using Mini Nutritional Assessment tool in Womberma district, West Gojjam Zone, Amhara Region, North West Ethiopia, 2020* [45]. Although about undernutrition among older adults in Ethiopia, the record was indexed with Infant, Newborn, likely because more records in MEDLINE about undernutrition in Ethiopia are concerned with infants (275 records) than older adults (26 records, both as of November 2023). Although records with these types of errors are relatively quick to exclude from search results, it is important to note that this type of indexing issue will become more frequent if algorithmically-indexed records form the foundation of future algorithmic indexing [46].

Finally, we found several examples of the MTIA improperly indexing a publication with the population being excluded, as illustrated by the article *Impact of the Controlling Nutritional Status (CONUT) score as a prognostic*

DOI: dx.doi.org/10.5195/jmla.2025.1936

factor for all-cause mortality in older patients without cancer receiving home medical care: hospital ward-based observational cohort study [47]. Despite expressly being about patients without cancer, the MTIA assigned the MeSH terms Neoplasms/Therapy, as well as Neoplasms as a major topic to its indexing.

LIMITATIONS AND FUTURE DIRECTIONS

Our study has significant limitations. First, we are not trained indexers. Our research did not seek to assess MTIA adherence to indexing guidelines, but rather sought to assess whether terms assigned by MTIA reflect the main concepts of an article and indicate its most important aspects. Consequently, our assessments of the representation of certain concepts may not have fully aligned with NLM indexing guidelines [48].

Another limitation is that publications from outside of our areas of experience were excluded. As medical librarians, our familiarity with MeSH is the result of years of literature searches supporting clinicians and knowledge synthesis projects [49-51] and providing instruction to health professionals. We found that we struggled to identify concepts from the titles and abstracts of records stemming from outside of clinical health; a difficulty also encountered in Liu and Wacholder [52]. We chose to exclude these articles because we were concerned that our inability to determine which parts of these publications would be relevant for a searcher would introduce inconsistency into our appraisal of MTIA performance.

We also note that we did not review screeners' individual assessments of the essential concepts within a record, nor did we check to make sure screeners agreed as to what those concepts were. Given the subjective nature of assigning subjects to a document, Golub's *A framework for evaluating automatic indexing or classification in the context of retrieval* [30] recommends that evaluations of indexing quality begin with expert consensus on all the relevant, appropriate subjects that should be assigned to it. Therefore, we recognize that our assessment of MTIA performance would be more solid if built on that additional foundation.

As previously noted, as of late April 2024, MTIA has been discontinued in favour of MTIX. In the March-April 2024 issue of its technical bulletin, the NLM highlights that the training data for MTIX is more recent, dating from 2007 through 2022, that MTIX considers the journal title where MTIA did not, and that MTIX can recognize the concept of "Hip Fractures" from interrupted and reordered phrases like "complex fractures and dislocations of the hip" [31]. The same publication asserts that the MTIX performance is comparable to that of MTIA at the level of precision (not erroneously applying terms) while improving on MTIA's recall (applying a greater number of terms which are not incorrect). However, MTIX, like MTIA, still does not access the full text of the publications it indexes. Moving forward, an obvious future direction would be the replication of this research with records indexed by MTIX.

We concur with Chen, Bullard and Giustini [26] that future research should ensure samples from a wide breadth of publications to assess the quality of indexing algorithm outputs in different fields. For example, Moore, Yaqub and Sampat [24] found that MTIA performed well for subject areas with specific terminology, such as diseases, and the NLM has indicated that chemicals and genes are priorities [53]. However, ensuring equitable quality of indexing across subjects will require ongoing research to evaluate indexing algorithm outputs in areas with innate lexical ambiguities like nursing, education or continuity of patient care.

We posit that future research should particularly scrutinize the accuracy of indexing for populations and its effects on retrieval. Some articles only identify populations of interest in their full text. As an example, Buono et al [54] has no indication of Black or African American people in the title or abstract, but on the basis of its full text, it has been indexed with the MeSH 'Black or African American'. As no MEDLINE platform presently permits a user to search within full text, MeSH indicating population groups, applied by human indexers based on the full text, constitute the only means for a searcher to find articles relevant to those groups. Further research could also appraise any disproportionate changes in numbers of records receiving MeSH for specific population subgroups or concepts.

Finally, when central concepts such as species, populations or publication types are omitted or inaccurately represented in indexing, search strategies or filters relying on MeSH-only queries may inadvertently overlook or exclude relevant publications. Filters designed and validated in a time when indexing was performed by humans, including such touchstones as the Cochrane Highly Sensitive Search Strategies [55], which has a MeSH-only line to exclude non-human animals, should be re-evaluated, as their performance may no longer be as reliable in this brave new world of inhuman indexing.

NOTE

Preliminary findings were presented the 2023 Canadian Health Library Association - Association des bibliothèques de la santé du Canada (CHLA-ABSC) conference; the authors also contributed commentary to the *Journal of European Association for Health Information and Libraries (JEAHIL)*.

ACKNOWLEDGEMENTS

We give special thanks to Dahlal Mohr-Elzeki, MA, MLIS (McGill University Health Centre) for her assistance with the manuscript. We are also grateful to Michele Datko, MS

DOI: dx.doi.org/10.5195/jmla.2025.1936

(ECRI Information Center), for a particularly significant example in the early stages of the project.

DATA AVAILABILITY STATEMENT

Data associated with this article are available in the Open Science Framework at https://osf.io/ckj3m/.

AUTHOR CONTRIBUTIONS

Alexandre Amar-Zifkin: conceptualization (equal), data curation, investigation (equal), methodology (equal), validation (equal), writing - original draft preparation and writing - review & editing (equal).
Taline Ekmekjian: investigation (equal), writing - original draft preparation and writing - review & editing (equal).
Virginie Paquet: investigation (equal), writing - original draft preparation and writing - review & editing (equal).
Tara Landry: conceptualization (equal), methodology (equal), project administration, resources, supervision, validation (equal), writing - original draft preparation and writing - review & editing (equal).

REFERENCES

1. International Organization for Standardization. Documentation--methods for examining documents, determining their subjects, and selecting indexing terms = Documentation--méthodes pour l'analyse des documents, la détermination de leur contenu et la sélection des termes d'indexation. [Geneva]: International Organization for Standardization; 1984.
2. Hjørland B. Indexing: Concepts and Theory. Knowledge Organization. 2018;45(7):609-39.
3. National Library of Medicine. Introduction to MeSH - Prefatory Remarks Bethesda, MD2023 [Available from: https://www.nlm.nih.gov/mesh/intro_preface.html#pref_rem].
4. Chang AA, Heskett KM, Davidson TM. Searching the Literature Using Medical Subject Headings versus Text Word with PubMed. Laryngoscope. 2006;116(2):336-40.
5. Jenuwine ES, Floyd JA. Comparison of Medical Subject Headings and text-word searches in MEDLINE to retrieve studies on sleep in healthy individuals. J Med Libr Assoc. 2004;92(3):349-53.
6. DeMars MM, Perruso C. MeSH and text-word search strategies: precision, recall, and their implications for library instruction. J Med Libr Assoc. 2022;110(1):23-33.
7. Swanberg SM, Dennison CC, Farrell A, Machel V, Marton C, O'Brien KK, et al. Instructional methods used by health sciences librarians to teach evidence-based practice (EBP): a systematic review. J Med Libr Assoc. 2016;104(3):197-208.
8. Neveol A, Shooshan SE, Humphrey SM, Mork JG, Aronson AR. A recent advance in the automatic indexing of the biomedical literature. J Biomed Inform. 2009;42(5):814-23.
9. Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf M-I, et al. Searching for and selecting studies. Cochrane Handbook for Systematic Reviews of Interventions 2019. p. 67-107.
10. Krithara A, Mork JG, Nentidis A, Paliouras G. The road from manual to automatic semantic indexing of biomedical literature: a 10 years journey. Front Res Metr Anal. 2023;8:1250930.
11. Mao Y, Lu Z. MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank. J Biomed Semantics. 2017;8(1):15.
12. National Library of Medicine. Citations Added to MEDLINE® by Fiscal Year 2022 2022 [Available from: https://www.nlm.nih.gov/bsd/stats/cit_added.html].
13. MEDLINE 2022 Initiative: Transition to Automated Indexing. The NLM technical bulletin. 2021;Nov-Dec(443):e5.
14. Mork JG, Jimeno-Yepes A, Aronson AR. The NLM Medical Text Indexer System for Indexing Biomedical Literature. BioASQ@ CLEF. 2013;1
15. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Stud Health Technol Inform. 2004;107(Pt 1):268-72.
16. National Library of Medicine. MTI Processing Flow White paper. 2006.
17. Dai S, You R, Lu Z, Huang X, Mamitsuka H, Zhu S. FullMeSH: improving large-scale MeSH indexing with full text. Bioinformatics. 2020;36(5):1533-41.
18. Mork J, Aronson A, Demner-Fushman D. 12 years on - Is the NLM medical text indexer still useful and relevant? J Biomed Semantics. 2017;8(1):8.
19. Minguet F, Salgado TM, van den Boogerd L, Fernandez-Llimos F. Quality of pharmacy-specific Medical Subject Headings (MeSH) assignment in pharmacy journals indexed in MEDLINE. Res Social Adm Pharm. 2015;11(5):686-95.
20. Tonin FS, Gmünder V, Bonetti AF, Mendes AM, Fernandez-Llimos F. Use of 'Pharmaceutical services' Medical Subject Headings (MeSH) in articles assessing pharmacists' interventions. Explor Res Clin Soc Pharm. 2022;7:100172.
21. Portaluppi F. Consistency and accuracy of the Medical Subject Headings® thesaurus for electronic indexing and retrieval of chronobiologic references. Chronobiol Int. 2007;24(6):1213-29.
22. Layton DM, Clarke M. Accuracy of Medical Subject Heading Indexing of Dental Survival Analyses. Int J Prosthodont. 2014;27(3):236-44.
23. Wilczynski NL, Haynes RB. Consistency and accuracy of indexing systematic review articles and meta-analyses in medline. Health Info Libr J. 2009;26(3):203-10.
24. Moore DAQ, Yaqub O, Sampat BN. Manual versus machine: How accurately does the Medical Text Indexer (MTI) classify different document types into disease areas? PLoS One. 2024;19(3):12.
25. Rae AR, Pritchard DO, Mork JG, Demner-Fushman D. Automatic MeSH Indexing: Revisiting the Subheading

DOI: dx.doi.org/10.5195/jmla.2025.1936

Attachment Problem. AMIA Annu Symp Proc. 2020;2020:1031-40.

26. Chen E, Bullard J, Giustini D. Automated indexing using NLM's Medical Text Indexer (MTI) compared to human indexing in Medline: a pilot study. J Med Libr Assoc. 2023;111(3):684-94.

27. Enoch J, Dickinson C, Potts J, Subramanian A. An exploratory study on support for caregivers of people with vision impairment in the UK. Ophthalmic Physiol Opt. 2022;42(4):858-71.

28. Lovay K, Barla J, Vasko J, Lendel A, Rakos M. Laparoscopic versus open elective right hemicolectomy with curative intent for colon adenocarcinoma. Rozhl Chir. 2022;100(12):584-91.

29. Lancaster FW. Indexing and abstracting in theory and practice. 3rd ed. Champaign (Ill.): University of Illinois, Graduate school of library and information Science; 2003.

30. Golub K, Soergel D, Buchanan G, Tudhope D, Lykke M, Hiom D. A framework for evaluating automatic indexing or classification in the context of retrieval. J Assoc Inf Sci Technol. 2015;67(1):3-16.

31. MTIX: the Next-Generation Algorithm for Automated Indexing of MEDLINE. The NLM technical bulletin. 2024;Mar-Apr(457):e4.

32. random.org. Random Integer Set Generator 2023 [Available from: https://www.random.org/integer-sets/].

33. Demas A. Blue as an orange. Lancet Neurol. 2023;22(3):207.

34. Ban JC, Hong JI, Wu YL. Mathematical analysis of topological and random m-order spread models. J Math Biol. 2023;86(3):40.

35. Zawar I, Ghosal S, Hantus S, Punia V. Indications for continuous electroencephalographic (cEEG) monitoring: What do they tell us? Epilepsy Res. 2023;190:107088.

36. Feng X, Ren L, Xiang Y, Xu Y. Development and validation of a nomogram for evaluating the incident risk of carotid atherosclerosis in patients with type 2 diabetes. Front Endocrinol (Lausanne). 2023;14:1131430.

37. Cho H, Houle SJD, Alsabbagh W. The trends and determinants of seasonal influenza vaccination after cardiovascular events in Canada: a repeated, pan-Canadian, cross-sectional study. Health Promot Chronic Dis Prev Can. 2023;43(2):87-97.

38. Ng CH, Tang ASP, Xiao J, Wong ZY, Yong JN, Fu CE, et al. Safety and tolerability of obeticholic acid in chronic liver disease: a pooled analysis of 1878 individuals. Hepatol Commun. 2023;7(3):e0005.

39. Krubiner CB, Tugendhaft A, DiStefano MJ, Barsdorf NW, Merritt MW, Goldstein SJ, et al. The Value of Explicit, Deliberative, and Context-Specified Ethics Analysis for Health Technology Assessment: Evidence From a Novel Approach Piloted in South Africa. Value Health Reg Issues. 2023;34:23-30.

40. Karim J, Gabrielli S, Mule P, Torabi B, Byrne A, De Schryver S, et al. Bridging knowledge gaps in paediatric chronic urticaria through a video-based educational tool. Clin Exp Dermatol. 2023;48(2):108-11.

41. Karanika S, Karantanos T, Carneiro H, Assoumou SA. Development and Validation of the HIV-CARDIO-PREDICT Score to Estimate the Risk of Cardiovascular Events in HIV-Infected Patients. Cells. 2023;12(4).

42. National Library of Medicine. 2021 MeSH Highlights 2021 [Available from: https://www.nlm.nih.gov/oet/ed/mesh/2021_mesh_highlights.html].

43. Guo MT, Gotz D, Wang Y. How Does Imperfect Automatic Indexing Affect Semantic Search Performance? Arxiv. 2023.

44. Wu B, Perin BI, Smith KD, Moshiri AS. Not All Cauliflowers Are HPV: Challenge. Am J Dermatopathol. 2023;45(3):e15.

45. Tadesse AD, Anto TG, Birhanu MY, Agedew E, Yimer B, Abejie AN. Prevalence of undernutrition and its associated factors among older adults using Mini Nutritional Assessment tool in Womberma district, West Gojjam Zone, Amhara Region, North West Ethiopia, 2020. PLoS One. 2023;18(2):e0274557.

46. Alemohammad S, Casco-Rodriguez J, Luzi L, Imtiaz Humayun A, Babaei H, LeJeune D, et al. Self-Consuming Generative Models Go MAD. 2023:31.

47. Miura H, Goto Y. Impact of the Controlling Nutritional Status (CONUT) score as a prognostic factor for all-cause mortality in older patients without cancer receiving home medical care: hospital ward-based observational cohort study. BMJ Open. 2023;13(2):e066121.

48. National Library of Medicine. MEDLINE Indexing Online Training Course: Indexing for MEDLINE 2015 [Available from: https://www.nlm.nih.gov/bsd/indexing/training/INT_030.html].

49. Maertens H, Madani A, Landry T, Vermassen F, Van Herzeele I, Aggarwal R. Systematic review of e-learning for surgical training. Br J Surg. 2016;103(11):1428-37.

50. Garant A, Guilbault C, Ekmekjian T, Greenwald Z, Murgoi P, Vuong T. Concomitant use of corticosteroids and immune checkpoint inhibitors in patients with hematologic or solid neoplasms: A systematic review. Crit Rev Oncol Hematol. 2017;120:86-92.

51. Fiore JF, Jr., El-Kefraoui C, Chay MA, Nguyen-Powanda P, Do U, Olleik G, et al. Opioid versus opioid-free analgesia after surgical discharge: a systematic review and meta-analysis of randomised trials. Lancet. 2022;399(10343):2280-93.

52. Liu Y-H, Wacholder N. Evaluating the impact of MeSH (Medical Subject Headings) terms on different types of searchers. Inf Process Manag. 2017;53(4):851-70.

53. National Library of Medicine. Minutes of the Board of Regents Meeting. 2023. p. 12.

54. Buono RJ, Bradfield JP, Wei Z, Sperling MR, Dlugos DJ, Privitera MD, et al. Genetic Variation in PADI6-PADI4 on 1p36.13 Is Associated with Common Forms of Human Generalized Epilepsy. Genes (Basel). 2021;12(9).

55. Cochrane Information Retrieval Methods Group. 4.4 Designing search strategies: Cochrane; 2023 [Version 6.4][Available from:

DOI: dx.doi.org/10.5195/jmla.2025.1936

https://training.cochrane.org/handbook/current/chapter-04#section-4-4-4].

AUTHORS' AFFILIATIONS

Alexandre Amar-Zifkin, MLIS, alexandre.amar-zifkin@umontreal.ca, https://orcid.org/0000-0002-1196-5209, Bibliothèque de la Santé, Université de Montréal, Montréal, Québec, Canada

Taline Ekmekjian, MLIS, taline.ekmekjian@phac-aspc.gc.ca, https://orcid.org/0000-0002-0914-0056, Public Health Agency of Canada library, Office of the Chief Science Officer, Public Health Agency of Canada, Ottawa, Ontario, Canada

Virginie Paquet, MIS, virginie.paquet@umontreal.ca, https://orcid.org/0000-0002-4931-134X, Bibliothèques de l'Université de Montréal, Montréal, Québec, Canada

Tara Landry, MA, MLIS, tara.landry@umontreal.ca, https://orcid.org/0000-0003-1930-9529, Bibliothèque de la Santé, Université de Montréal, Montréal, Québec, Canada

Received February 2024; accepted October 2024

ISSN 1558-9439 (Online)

ORIGINAL INVESTIGATION
DOI: dx.doi.org/10.5195/jmla.2025.1591

What's beyond the core? Database coverage in qualitative information reveal

Jennifer Horton; David Kaunelis; Danielle Rabb; Andrea Smith

See end of article for authors' affiliations.

Objective: This study investigates the effectiveness of bibliographic databases to retrieve qualitative studies for use in systematic and rapid reviews in Health Technology Assessment (HTA) research. Qualitative research is becoming more prevalent in reviews and health technology assessment, but standardized search methodologies—particularly regarding database selection—are still in development.

Methods: To determine how commonly used databases (MEDLINE, CINAHL, PsycINFO, Scopus, and Web of Science) perform, a comprehensive list of relevant journal titles was compiled using InCites Journal Citation Reports and validated by qualitative researchers at Canada's Drug Agency (formerly CADTH). This list was used to evaluate the qualitative holdings of each database, by calculating the percentage of total titles held in each database, as well as the number of unique titles per database.

Results: While publications on qualitative search methodology generally recommend subject-specific health databases including MEDLINE, CINAHL, and PsycINFO, this study found that multidisciplinary citation indexes Scopus and Web of Science Core Collection not only had the highest percentages of total titles held, but also a higher number of unique titles.

Conclusions: These indexes have potential utility in qualitative search strategies, if only for supplementing other database searches with unique records. This potential was investigated via tests on qualitative rapid review search strategies translated to Scopus to determine how the index may contribute relevant literature.

Keywords: Informative retrieval; Qualitative research; Evidence synthesis; Database selection

See end of article for supplemental content.

INTRODUCTION

Qualitative evidence synthesis approaches are becoming more prevalent in health technology assessments (HTAs). Studies which employ qualitative research methods are useful when considering patient experience and preferences as well as the observations of clinical experts. Qualitative evidence synthesis (QES) is a set of methodologies used to conduct systematic evidence synthesis of primary qualitative research [1, 2]. In HTA, an evidence-based field where study design, results reporting, and review protocols are standardized, researchers and reviewers may struggle with including qualitative perspectives—which by their very nature, must be analyzed, synthesized, and critically appraised differently than clinical or economic information typically addressed in health technology assessments [3]. Though the methodological differences between reviews of effectiveness or cost-effectiveness and qualitative reviews may seem at odds with each other, more recent literature takes a reconciliatory approach, such as Booth's 2018 article on the "dual heritage" of QES [3]. Here, Booth argues that QES draws on methodologies from primary qualitative research as well as knowledge synthesis of clinical primary information, thus allowing more opportunity to utilize a variety of methodological approaches for analysis.

This study focuses on the bibliographic databases that can be used to retrieve qualitative research in the context of rapid reviews. Such reviews are carried out in a shorter time frame than systematic or scoping qualitative reviews and are often conducted in relation to a specific decision-making need and often answer more focused and narrower research questions [4, 5]. The characteristics of rapid QES inform information retrieval methodology which includes less exhaustive search strategies and a focus on a manageable number of results for an expedited timeline. Much has been written in the past ten years on how to retrieve studies for use in systematic and rapid

DOI: dx.doi.org/10.5195/jmla.2025.1591

QES, and while there is no single standard method, common practices can be pieced together [2, 6-13], which share at least three essential components—which databases to search, which search filters (if any) to employ, and how to screen or select relevant sources. This study focuses on database selection.

Just as in quantitative study searches, the databases selected have a considerable impact on the yield and relevance of qualitative results retrieved [14, 15]. Yet, prominent resources on conducting qualitative systematic reviews—such as the Joanna Briggs Institute Reviewer's Manual and the Cochrane Handbook for Systematic Reviews—do not recommend specific databases* [8, 16, 17]. The Centre for Reviews and Dissemination's Guidance for Undertaking Reviews in Health Care points to recently added qualitative subject headings in MEDLINE and CINAHL but does not make an explicit recommendation to use these databases [18].

Recent literature takes up the task from which a list of databases can be compiled. The two most mentioned are MEDLINE and the Cumulative Index to Nursing and Allied Health Literature (CINAHL) [6, 11]. MEDLINE provides the most comprehensive collection of health science research and is used heavily in both quantitative and qualitative searches [2, 14]. CINAHL contains a higher percentage of qualitative research than MEDLINE, with 4% to 5% compared to 1% of total database holdings [11]. Embase and PsycINFO are also recommended. While holding unique results, Embase was employed less often in literature on qualitative searches. According to a study by Subirana et al and cited by Booth in 2016, Embase retrieved minimal unique results [2, 19]. Additionally, Frandsen et al verifies these findings and in their recommendations for database selection with maximum recall, Embase is not included in any combination [7]. Due to the authors' focus on rapid QES which necessitates a compromise on number of databases searched and yield of results that are feasible to screen in shorter amounts of time, Embase will not be included in our analysis [14]. Like CINAHL, the subject-specific nature of PsycINFO limits its utility. Subject-specificity is not always a limitation, however, an HTA organization most likely would not consider subscriptions to subject specific bibliographic databases to be particularly cost-effective considering the financial costs per use. The last databases on this list are the multidisciplinary citation indexes Scopus and Web of Science Core Collection (WoS). Until recently, as discussed more fully in Frandsen et al's 2019 article, these indexes had not been focused on or utilized as frequently as the other databases mentioned [7, 14].

The core set of databases used for other health-based systematic reviews—MEDLINE, CINAHL, and PsycINFO—is perhaps the most logical place to start searching due to their subject-specific nature, and these databases are the most likely to be already available to researchers, particularly in the HTA context. MEDLINE and CINAHL contain qualitative subject headings, and while indexing is strong for qualitative information in CINAHL, researchers will often also employ qualitative study filters to retrieve the most relevant results.

Database selection is especially pertinent in rapid QES to ensure that the retrieval of as many relevant studies is possible and feasible, focusing on a breadth and richness of differing perspectives on the same question [2, 20]. As discussed previously, recommendations for a core set of databases to search are sparse. This study aims to investigate the utility of specific databases and citation indexes to identify and balance the most—and most relevant—qualitative primary studies that are realistically manageable within the context of rapid reviews. This study can inform the selection of a core set of databases that will allow researchers to maximize the number of unique results and avoid searching more resources with fewer returns.

There is a case to be made for resources like Scopus and WoS to be included in a core set of databases to search for QES. Frandsen et al 2019 article gives strong evidence for use of Scopus, but this finding must be taken in context of their study. Tests run by Frandsen et al in Scopus did not exclude records also indexed in MEDLINE and Embase. This operation inflates the number of records retrieved by Scopus. It should be noted that running a search for MEDLINE and Embase records solely in Scopus is risky. The lack of hierarchical and standardized subject headings in Scopus that are available in MEDLINE (MeSH) and Embase (EMTREE) make searching these databases in Scopus less precise. This indicates that searches must still be run in other platforms such as Ovid which host MEDLINE or Embase to retrieve the best quality results for those databases. Frandsen et al also chose databases to analyze retroactively, based on those indicated in reports chosen. As a result, the percentage of studies retrieved by each database is again inflated, as these were the only databases searched for the reports in the first place. Frandsen's findings on Scopus nevertheless raise important questions regarding multidisciplinary citation indexes, which this present study further explores in its latter part through test searches [7].

Given that database selection is key to ensuring breadth in QES, this study aims to evaluate databases based on their

* Though not making a recommendation as to where to conduct a search, the Cochrane Handbook does point to qualitative search filters for MEDLINE, EMBASE, CINAHL, and PsycINFO (Cochrane 2011, 20.3.2.1).

DOI: dx.doi.org/10.5195/jmla.2025.1591

holdings of qualitative information. These findings will help clarify which databases are useful and efficient when developing search strategies for QES, particularly within the context of HTA wherein researchers or information specialists may not have access to as many databases as are available at a research-intensive university. The second part of this study explores multidisciplinary citation indexes to determine how useful they may be at retrieving qualitative information in practice.

METHODS

This study takes a modified approach to database evaluation by comparing holdings of a predetermined list of relevant journal titles. Previous studies on literature mapping for various disciplines employ similar methods to comprehensively analyze core journals to extract pertinent titles for disciplines such as social work and physical therapy [21, 22]. Similar practices are used to compare subject holdings across databases [23-25]. In the case of qualitative research related to the health sciences, a disciplinary mapping of the literature is more complex, as these studies can appear across a variety of discipline-specific publications. For this reason, the authors did not adhere to the literature mapping protocol laid out by the Medical Library Association [26]. Instead, we assessed the selected databases using a set of relevant journals to determine coverage of the topic area. After assessing the holdings of different databases, the authors sought to explore the performance of multidisciplinary citation indexes to retrieve qualitative studies. We then conducted a series of tests in Scopus to determine how searching this multidisciplinary index contributed to the overall results of a series of qualitative rapid reviews conducted by Canada's Drug Agency (formerly CADTH), a Canadian HTA agency.

The first part of the study began by compiling a list of journal titles based on a shortlist of frequently consulted titles from qualitative researchers at Canada's Drug Agency (see supplementary materials). The titles on this shortlist were searched in Clarivate Analytics InCites Journal Citation Reports (JCR) to determine which subject categories they fell into. All journal titles in the following categories were exported for analysis—Anthropology; Cultural Studies; Health Policy & Services; Social Sciences; Biomedical; and Social Issues. Categories related to health sciences but lacking a social science disciplinary aspect were excluded, as these journals are the focus of the core set of health science databases. A total of 286 titles were exported from JCR into a spreadsheet for analysis. This set was then sent to the qualitative team at Canada's Drug Agency for validation, supplementing, and secondary screening of the titles for relevancy which brought the final list to 191 titles. Using this list, the authors consulted Ulrich's Web to determine where each title was indexed. Attention was paid specifically to commonly used databases for HTA—including MEDLINE, CINAHL, PsycINFO—as well as multidisciplinary databases Scopus and WoS. Since the set of journal titles came from JCR, which draws on information from WoS and its holdings, the authors' study results on WoS are skewed. However, by assessing other databases, JCR and WoS holdings are externally validated, and the set still provides a comprehensive set with which to test these other databases. Additionally, starting with WoS holdings allowed the authors to utilize a large set of titles that is interdisciplinary, geographically diverse, and has a minimum of predatory titles. This set was also just a starting point to present to qualitative researchers to validate externally. It should also be noted here that holdings in Ulrich's Web were recorded in terms of presence or no presence and did not consider date ranges of these holdings in each database.

In addition to evaluating databases based on journals indexed, a second component of the study goes further in testing to determine how Scopus performed and contributed to previously run searches for nine published qualitative rapid reviews [27-35]. Scopus was chosen over WoS because it is the multidisciplinary databases subscribed to by the authors, and because the test set of journal titles came from JCR, a Clarivate product informed by the holdings of WoS. The authors chose nine qualitative rapid reviews to assess due to readily available information such as comprehensive search method documentation and existing EndNote libraries, which were easily accessed in-house. Search strategies from the nine rapid reviews were translated into Scopus and combined with a translation of the Canada's Drug Agency qualitative study filter <https://searchfilters.cadth.ca>. Search strings were directly translated wherever possible, with the exception of MeSH headings. If heading words or phrases were not also covered in title and abstract queries, they were added. Scopus has a very general controlled vocabulary based on journal subject categories, with headings like Social Science and Medicine which are too vague to include in a search. Though Scopus will list MeSH and EMTREE headings for articles pulled from MEDLINE and Embase, users cannot search for these terms as controlled vocabulary, only as keywords. Once the searches were translated, they were run in Scopus, with MEDLINE and Embase results excluded. These database results were excluded from Scopus searches so that the authors could evaluate Scopus on its own. Also as previously discussed, it is not generally good practice to search MEDLINE or Embase within Scopus for reviews due to less sophisticated search functionality offered on Scopus versus other platforms. Date and language limits were also applied when applicable. Search results were exported to EndNote and compared with existing libraries of literature search results for each rapid review. The authors manually deduplicated in EndNote to ensure that all Scopus results were unique. All unique Scopus results were then screened by one author who is an experienced qualitative researcher. In the first level of screening, titles

DOI: dx.doi.org/10.5195/jmla.2025.1591

and abstracts were reviewed and potentially relevant full-text articles were retrieved and assessed for inclusion. The final selection of articles was based on the inclusion criteria in the published rapid reviews.

RESULTS

Database Assessment via Core Journal Holdings

The set of 191 journal titles was compiled from JCR and validated by qualitative researchers at Canada's Drug Agency. This list was analyzed to determine where each title is indexed, using Ulrich's Web. Information collected from Ulrich's Web on where each journal title is indexed was further analyzed to determine the percentage of titles covered in each database, as well as the percentage of titles unique to each database. Not all 191 journals on the list were indexed in the databases studied, as indicated in Table 1. Multidisciplinary databases have the highest percentage of total holdings, with WoS at 91% and Scopus at 82%. MEDLINE had the second highest percentage of total holdings, followed by more subject-specific databases CINAHL (47%) and PsycINFO (38%).

Table 1 Percentage of Journal Titles per Database
*Web of Science Core Collection contains Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, and Emerging Sources Citation Index

Database	Ratio of Total Titles	Percentage of Total Titles	Ratio of Unique Titles	Percentage of Unique Titles
MEDLINE	98/191	51%	0/191	0%
PsycINFO	73/191	38%	0/191	0%
CINAHL	91/191	48%	0/191	0%
Scopus	157/191	82%	1/191	0.5%
Web of Science*	175/191	92%	6/191	3%
Not indexed	13/191	7%	13/191	7%

While these results do indicate that both Scopus and WoS retrieve unique results, it is unlikely that a search approach would employ both multidisciplinary databases. There is a considerable amount of overlap between the two, which skews the percentage of unique titles for each in Table 1. More calculations were done to determine how many unique results each multidisciplinary database would yield compared to the core set of health science databases when the other was excluded from the data set (Table 2).

Table 2 Percentage of Unique Titles (with exclusions)

Database	Ratio of Unique Titles (with exclusions)	Percentage of Unique Titles (with exclusions)
Scopus (excluding Web of Science)	49/191	26%
Web of Science (excluding Scopus)	55/191	29%

This set illustrates a higher percentage of unique holdings when compared only with core and subject-specific health databases, indicating a considerable possible benefit for searching a multidisciplinary database in addition to common HTA resources.

Running Searches: Scopus Assessment

To better understand the potential benefits of employing a multidisciplinary database, searches from previous CADTH qualitative rapid reviews were translated and run in Scopus to determine the number of unique results for each (Table 3). From a purely quantitative perspective, Scopus retrieves a significant number of unique results, which have the potential to facilitate the breadth of perspectives that is important in QES. Adding more results, however, does not necessarily lead to additional relevant studies. To further assess these results, qualitative researchers at Canada's Drug Agency screened citations from Scopus for inclusion for each rapid review topic. Exact results are included below (Table 4).

DOI: dx.doi.org/10.5195/jmla.2025.1591

Table 3 Scopus Citation Comparison on CADTH Rapid Qualitative Reviews

Qualitative Rapid Review	Existing EndNote Records	Original Databases Searched	New Scopus Results	Percentage of New Results
Engaging with History Taking for Adverse Childhood Experiences in Care: A Rapid Qualitative Review [27]	1596	MEDLINE CINAHL PsycINFO	300	19%
Prescription Drug Monitoring Programs: A Rapid Qualitative Review [28]	89	PubMed	6	7%
Gene Expression Profiling Tests for Breast Cancer: A Rapid Qualitative Review [29]	181	MEDLINE CINAHL	24	13%
Rural Breast Cancer Surgery Programs: A Rapid Qualitative Review [30]	443	MEDLINE CINAHL	161	36%
Prostatectomy for People with Prostate Cancer: A Rapid Qualitative Review [31]	839	MEDLINE CINAHL	15	2%
Biopsy for Adults with Suspected Skin Cancer: A Rapid Qualitative Review [32]	602	MEDLINE CINAHL	10	2%
Screening and Diagnostic Services for People at Risk for Breast Cancer: A Rapid Qualitative Review [33]	995	PubMed Cochrane	54	5%
Experiences with and Expectations of Robotic Surgical Systems: A Rapid Qualitative Review [34]	1031	MEDLINE PsycINFO Scopus	283	27%
Point-of-Care Testing of International Normalized Ratios for People on Oral Anticoagulants: A Rapid Qualitative Review [35]	426	MEDLINE Embase Scopus	65	15%

Table 4 Relevance of Scopus Results in qualitative rapid reviews

Qualitative Rapid Review	Total Results*	Included Studies*	Percentage of Studies Included	Total Scopus Results	Scopus Results Selected for Inclusion	Percentage of Studies Included (Scopus)
Engaging with History Taking for Adverse Childhood Experiences in Care: A Rapid Qualitative Review [27]	1596	6	0.38%	300	1	0.33%
Prescription Drug Monitoring Programs: A Rapid Qualitative Review [28]	89	18	20.22%	6	0	0%
Gene Expression Profiling Tests for Breast Cancer: A Rapid Qualitative Review [29]	181	11	6.08%	24	0	0%

DOI: dx.doi.org/10.5195/jmla.2025.1591

Rural Breast Cancer Surgery Programs: A Rapid Qualitative Review [30]	443	12	2.71%	161	0	0%
Prostatectomy for People with Prostate Cancer: A Rapid Qualitative Review [31]	854	38	4.45%	15	0	0%
Biopsy for Adults with Suspected Skin Cancer: A Rapid Qualitative Review [32]	612	12	1.96%	10	0	0%
Screening and Diagnostic Services for People at Risk for Breast Cancer: A Rapid Qualitative Review [33]	1049	12	1.14%	54	0	0%
Experiences with and Expectations of Robotic Surgical Systems: A Rapid Qualitative Review [34]	1031	14	1%	283	0	0%
Point-of-Care Testing of International Normalized Ratios for People on Oral Anticoagulants: A Rapid Qualitative Review [35]	426	5	1.2%	65	3	0.7%

* According to report PRISMA flowcharts

Such a small set of reports tested in Scopus can only provide a limited perspective. From this test, however, just one study was selected for inclusion from the Scopus results for the first report "Engaging with History Taking for Adverse Childhood Experiences in Care" and three selected from "Point-of-Care Testing of International Normalized Ratios for People on Oral Anticoagulants" [27, 35]. Comparing total Scopus results to included Scopus studies with that of the original report results—which searched only core health science databases—indicates that this ratio can vary depending on the research question, from as high as 20% to as little as 0.38%. Percentage of studies included between the original reports and the accompanying Scopus tests are not wildly different in most cases, which may show that the ratio of included studies to the total number of results is more question-dependent as opposed to databases-specific. As such, this small sample of reports analyzed does not represent a large breadth of research topic areas. One can imagine that in other topic areas, utilizing multidisciplinary citation indexes could be more useful.

DISCUSSION

As observed from the first part of this study, databases CINAHL and PsycINFO did not include any unique holdings based on the list of journals searched for. However, their controlled vocabulary and indexing are unique, and it is likely one would retrieve unique citations that may be relevant for QES. Thus, a search in CINAHL can retrieve unique results. This validates the methodological practice of searching multiple databases that have similar holdings, as results may differ based on the search strategy which may include different subject headings, holding completeness of certain titles (date ranges of title held), and search functionality specificities.

Scopus and Web of Science have a high number of unique journal titles that would not be searched at all if only adhering to core health science databases. It also important to note that Scopus searches tested do not add significantly to the original report results—meaning that the total number of results would still be a reasonable amount for a reviewer to screen. Scopus is proven here for retrieval of unique results, but more tests must be done to assess the quality of results retrieved by a multidisciplinary citation index such as Scopus, and changes to search strategy (aside from a direct translation) may be necessary.

This research is subject to several limitations. The authors' decision to exclude Embase from study was informed by literature on the topic, but there is no consensus on whether searching Embase for qualitative literature is beneficial. Thus, this study cannot recommend or discourage use of Embase, and further research must be conducted to evaluate this database.

Journal titles were chosen based on subjects in JCR and supplemented by qualitative researchers at Canada's Drug Agency, but the list used is not an exhaustive one. JCR, the product used to compile an initial list of journal titles, solely includes holdings of WoS and does not take into account additional holdings of Scopus. This resulted in Scopus's holdings being underreported by the authors. Further research must be done using other journal lists to better determine the breadth of qualitative titles in Scopus. The authors' decision to use Ulrich's Web to evaluate database holdings did not account for date ranges of each journal held in each database. This may imply complete journal coverage in a given database, but actual holdings were not verified in this study. A more accurate approach

DOI: dx.doi.org/10.5195/jmla.2025.1591

may have been to have searched the journal title list of each database, however additional research may be conducted to more accurately verify database holdings. Scopus test cases illustrate that implementation is most important, but such concerns as journal date coverage is built into common practice for systematic searching. Further investigation into date range discrepancies between databases could be conducted, but it was out of scope for this study.

In addition to being limited to the researchers' experience with and familiarity of journal titles, it is difficult to determine exactly where qualitative research is being published, especially because more and more qualitative studies are being included in otherwise clinical publications. Qualitative searching and the volume of literature retrieved are heavily dependent on the research topic and question, and while this study's sample of nine qualitative rapid responses provides some insight into how multidisciplinary databases contribute to a search, it is far from being comprehensive given the limited number of research topics covered by these reviews.

Additionally, database selection is just one component to information retrieval. Other factors such as search strategy and filter selection have an impact on the volume and relevancy of literature results. This study is also biased towards the needs—and resources—of the organization at which it was conducted. The research practice and goals of Canada's Drug Agency are characteristic of the HTA field but its individual mission and the resources available inform the purpose and limitations of our study.

This study focused on rapid reviews specifically and did not address other more comprehensive forms of knowledge synthesis. Traditional systematic or scoping reviews have different methodological requirements and may necessitate use of more databases than those studied here. Testing database search capabilities in those contexts therefore may produce different results than the rapid reviews studied here. Research questions and search strategies for rapid reviews are typically more focused than those for a systematic review, and this focus may have limited the number of results retrieved in the authors' test of Scopus. In other words, a broader search question may yield more potentially relevant results.

Lastly, the decision to directly translate search strings from MEDLINE to Scopus for the evaluation of a multidisciplinary citation index influenced the number and quality of results retrieved. MeSH terms from the original searches could not be searched or adequately translated due to the broad nature of Scopus's Subject Area controlled vocabulary.

The findings presented here further our understanding of the utility of various databases in QES and simultaneously raise more questions related to qualitative search methodology. Considering the percentage of unique journal titles indexed in Scopus and WoS, what place do these resources have in regard to QES information retrieval methods? If multidisciplinary citation indexes were included along with the core health science databases, in what ways would search strategies need to change to yield results of the best quality—and of reasonable quantity? Might Scopus journal subject categories be employed to exclude out-of-scope disciplines such as chemistry, engineering, and physics?

These questions help determine next steps for further study. To gain more accurate insights into Scopus performance in QES, information specialists at Canada's Drug Agency will continue to search the database on a trial, case-by-case basis. Information specialists working on projects with a qualitative focus or component (such as rapid and systematic reviews) will search Scopus in addition to core health science databases. Number of Scopus results retrieved, and studies to be included will be documented in a similar manner to the tests done on previous rapid reviews. In addition, there will be opportunities to consult with the information specialists and qualitative reviewers on their reflections and lessons learned to adjust and improve Scopus search practice.

CONCLUSION

Along with the set of core health science databases (MEDLINE/PubMed, PsycINFO, CINAHL), it can be beneficial to include Scopus or Web of Science as a supplemental source of qualitative information. These multidisciplinary indexes contain unique journals which publish relevant study types—studies which may otherwise be missed if only searching core health science databases. At the same time, the authors acknowledge the potential limitations of searching in Scopus or Web of Science. Search functionality in these indexes is not as complex or controlled as searching in health science databases, particularly due to the lack of an adequate controlled vocabulary. These databases are also costly to license, which may make them inaccessible to HTA agencies or other organizations. One must also note that the yield of relevant information from Scopus or Web of Science is question-specific. Though the ideal standard for evidence syntheses is to locate and consider all possible sources of information, research teams must contend with how they can at once be most efficient and most thorough, balancing precision with sensitivity given resource constraints and shortened time frames. Further research should be conducted to determine just how useful the addition of a multidisciplinary index like Scopus or Web of Science may be to rapid qualitative evidence synthesis projects—as well as search or translation strategies which best fit these indexes—though this study provides evidence that such indexes show promise in the field of QES.

DOI: dx.doi.org/10.5195/jmla.2025.1591

ACKNOWLEDGMENTS

We thank Elijah Herington for sharing his expertise and guidance in the conduct of this study. We also thank Francesca Brundisini for providing feedback on a draft.

DATA AVAILABILITY STATEMENT

Data and search strategies associated with this article are available in the Open Science Framework at https://osf.io/dj7gf/

AUTHOR CONTRIBUTIONS

Jennifer Horton: conceptualization, investigation, methodology, writing - original draft. David Kaunelis: conceptualization, writing - review & editing.Danielle Rabb: conceptualization, writing - review & editing. Andrea Smith: conceptualization, writing - review & editing.

REFERENCES

1. Flemming K, Booth A, Garside R, Tunçalp Ö, Noyes J. Qualitative evidence synthesis for complex interventions and guideline development: clarification of the purpose, designs and relevant methods. BMJ Global Health. 2019 4(Suppl 1):e000882. http://gh.bmj.com/content/4/Suppl_1/e000882.abstract.
2. Booth A. Searching for qualitative research for inclusion in systematic reviews: a structured methodological review. Systematic Reviews. 2016 2016/05/04;5(1):74. https://doi.org/10.1186/s13643-016-0249-x.
3. Booth A. Harnessing Energies, Resolving Tensions: Acknowledging a Dual Heritage for Qualitative Evidence Synthesis. Qualitative Health Research. 2018 2019/01/01;29(1):18-31. https://doi.org/10.1177/1049732318808247.
4. Kelly SE, Moher D, Clifford TJ. DEFINING RAPID REVIEWS: A MODIFIED DELPHI CONSENSUS APPROACH. Int J Technol Assess Health Care. 2016 Jan;32(4):265-75. https://pubmed.ncbi.nlm.nih.gov/27724987/.
5. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information & Libraries Journal. 2009 2009/06/01;26(2):91-108. https://doi.org/10.1111/j.1471-1842.2009.00848.x.
6. Campbell F, Weeks L, Booth A, Kaunelis D, Smith A. A Scoping Review Found Increasing Examples of Rapid Qualitative Evidence Syntheses and no Methodological Guidance. Journal of clinical epidemiology. 2019 Jun 20. https://pubmed.ncbi.nlm.nih.gov/31229582/.
7. Frandsen TF, Gildberg FA, Tingleff EB. Searching for qualitative health research required several databases and alternative search strategies: a study of coverage in bibliographic databases. Journal of clinical epidemiology. 2019 114:118-24. https://doi.org/10.1016/j.jclinepi.2019.06.013.
8. Booth A, Cooper C, Dracup N. Qualitative Research. SuRe Info. 2024. https://sites.google.com/york.ac.uk/sureinfo/home/qualitative-research.
9. Booth A. Qualitative Evidence Synthesis. In: Karen M. Facey, Helle Ploug Hansen, and Ann N. V. Single, editors. Patient Involvement in Health Technology Assessment. Singapore: Springer Singapore; 2017. p. 187-99.
10. Selva A, Solà I, Zhang Y, Pardo-Hernandez H, Haynes RB, Martínez García L, Navarro T, Schünemann H, Alonso-Coello P. Development and use of a content search strategy for retrieving studies on patients' views and preferences. Health and Quality of Life Outcomes. 2017 2017/08/30;15(1):126. https://doi.org/10.1186/s12955-017-0698-5.
11. DeJean D, Giacomini M, Simeonov D, Smith A. Finding Qualitative Research Evidence for Health Technology Assessment. Qualitative Health Research. 2016 26(10):1307-17. https://journals.sagepub.com/doi/abs/10.1177/1049732316644429.
12. Pearson M, Moxham T, Ashton K. Effectiveness of Search Strategies for Qualitative Research About Barriers and Facilitators of Program Delivery. Evaluation & the Health Professions. 2011 2011/09/01;34(3):297-308. https://doi.org/10.1177/0163278710388029.
13. Wagner M, Rosumeck S, Küffmeier C, Döring K, Euler U. A validation study revealed differences in design and performance of MEDLINE search filters for qualitative research. Journal of clinical epidemiology. 2020 Apr;120:17-24. https://pubmed.ncbi.nlm.nih.gov/31862229/.
14. Cooper C, Booth A, Varley-Campbell J, Britten N, Garside R. Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies. BMC Medical Research Methodology. 2018 2018/08/14;18(1):85. https://doi.org/10.1186/s12874-018-0545-3.
15. Hartling L, Featherstone R, Nuspl M, Shave K, Dryden DM, Vandermeer B. The contribution of databases to the results of systematic reviews: a cross-sectional study. BMC Medical Research Methodology. 2016 2016/09/26;16(1):127. https://doi.org/10.1186/s12874-016-0232-1.
16. Lockwood CP, Munn Z, Rittenmeyer L, Salmond S, Bjerrum M, Loveday H, Carrier J, Stannard D. Chapter 2: Systematic reviews of qualitative evidence. 2017. In: Joanna Briggs Institute Reviewer's Manual [Internet]. Adelaide (Australia): Joanna Briggs Institute. Available from: https://jbi-global-wiki.refined.site/space/MANUAL/4688637/Chapter+2%3A+Systematic+reviews+of+qualitative+evidence.
17. Lefebvre C, Glanville J, Briscoe S, Featherstone R, Littlewood A, Marshall C, Metzendorf M-I, Noel-Storr A, Paynter R, Rader T, Thomas J, Wieland LS. Chapter 4: Searching for and selecting studies. 2019. In: Cochrane Handbook for Systematic Reviews of Interventions [Internet]. London (UK): Cochrane. Available from: https://training.cochrane.org/handbook/current/chapter-04.
18. Systematic Reviews: CRD's guidance for undertaking reviews in health care. York (UK): Centre for Reviews and

DOI: dx.doi.org/10.5195/jmla.2025.1591

Dissemination, University of York; 2009. Available from: https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf.

19. Subirana M, Solá I, Garcia JM, Gich I, Urrútia G. A nursing qualitative systematic review required MEDLINE and CINAHL for study identification. Journal of clinical epidemiology. 2005 2005/01/01/;58(1):20-5. https://www.sciencedirect.com/science/article/pii/S089543560400167 2.

20. Brunton G, Stansfield C, Caird J, Thomas J. Finding Relevant Studies. In: Gough DO and Thomas J, editors. An Introduction to Systematic Reviews. London (UK): SAGE; 2012. p. 107-34.

21. Peace J, Brennan PF. MetaMapping the nursing procedure manual. AMIA Annu Symp Proc. 2006 2006:1060. https://pubmed.ncbi.nlm.nih.gov/17238679/.

22. Fell DW, Burnham JF, Buchanan MJ, Horchen HA, Scherr JA. Mapping the core journals of the physical therapy literature. J Med Libr Assoc. 2011 99(3):202-7. https://pubmed.ncbi.nlm.nih.gov/21753912.

23. Mayr P, Walter A-K. Studying Journal Coverage in Google Scholar. Journal of Library Administration. 2008 2008/09/30;47(1-2):81-99. https://doi.org/10.1080/01930820802110894.

24. Salisbury L. Web of Science and Scopus: A comparative review of content and searching capabilities. The Charleston Advisor. 2009 July:5-18. https://charleston.publisher.ingentaconnect.com/content/charleston/chadv/2009/00000011/00000001/art00005#.

25. Mongeon P, Paul-Hus A. The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics. 2016 January 01;106(1):213-28. https://doi.org/10.1007/s11192-015-1765-5.

26. Allen MP, Jacobs SK, Levy JR. Mapping the literature of nursing: 1996-2000. J Med Libr Assoc. 2006 94(2):206-20. https://pubmed.ncbi.nlm.nih.gov/16636714.

27. Engaging with History Taking for Adverse Childhood Experiences in Care: A Rapid Qualitative Review. Ottawa (ON): CADTH; 2018. Available from: https://www.cda-amc.ca/sites/default/files/rr/2019/RC1045%20History%20Taking%20ACE%20Final.pdf.

28. Prescription Drug Monitoring Programs: A Rapid Qualitative Review. Ottawa (ON): CADTH; 2019. Available from: https://www.cda-amc.ca/prescription-drug-monitoring-programs-rapid-qualitative-review.

29. Gene Expression Profiling Tests for Breast Cancer: A Rapid Qualitative Review. Ottawa (ON): CADTH; 2019. Available from: https://www.cda-amc.ca/sites/default/files/rr/2019/RC1099%20Gene%20Profiling%20BrCa%20Final.pdf.

30. Rural Breast Cancer Surgery Programs: A Rapid Qualitative Review Ottawa (ON): CADTH; 2019. Available from: https://www.cda-amc.ca/sites/default/files/rr/2019/RC1104%20Rural%20Breast%20Cancer%20Surgery%20Final.pdf.

31. Prostatectomy for People with Prostate Cancer: A Rapid Qualitative Review. Ottawa (ON): CADTH; 2019. Available from: https://www.cda-amc.ca/prostatectomy-people-prostate-cancer-rapid-qualitative-review.

32. Biopsy for Adults with Suspected Skin Cancer: A Rapid Qualitative Review. Ottawa (ON): CADTH; 2019. Available from: https://www.cda-amc.ca/biopsy-adults-suspected-skin-cancer-rapid-qualitative-review.

33. Screening and Diagnostic Services for People at Risk for Breast Cancer: A Rapid Qualitative Review. Ottawa (ON): CADTH; 2019. Available from: https://www.ncbi.nlm.nih.gov/books/NBK551875/.

34. Experiences with and Expectations of Robotic Surgical Systems: A Rapid Qualitative Review. Ottawa (ON): CADTH; 2020. Available from: https://www.cda-amc.ca/experiences-and-expectations-robotic-surgical-systems-rapid-qualitative-review.

35. Point-of-Care Testing of International Normalized Ratios for People on Oral Anticoagulants: A Rapid Qualitative Review. Ottawa (ON): CADTH; 2021. Available from: https://www.cda-amc.ca/point-care-testing-international-normalized-ratios-people-oral-anticoagulants-rapid-qualitative.

SUPPLEMENTAL FILES

- **Appendix A**: Supplemental Data

AUTHORS' AFFILIATIONS

Jennifer Horton, jhorton@cap.org, https://orcid.org/0000-0002-8213-9297, Medical Librarian, College of American Pathologists, Northfield, Illinois, United States

David Kaunelis, David.Kaunelis@cda-amc.ca, Methods Specialist, CDA-AMC (Canada's Drug Agency), Ottawa, Ontario, Canada

Danielle Rabb, Danielle.Rabb@cda-amc.ca, Director-Research Information Services, CDA-AMC (Canada's Drug Agency), Ottawa, Ontario, Canada

Andrea Smith, Andrea.Smith@cda-amc.ca, Lead-Strategic Projects, CDA-AMC (Canada's Drug Agency), Ottawa, Ontario, Canada

Received May 2022; accepted October 2024

ISSN 1558-9439 (Online)

ORIGINAL INVESTIGATION
DOI: dx.doi.org/10.5195/jmla.2025.1972

Filtering failure: the impact of automated indexing in Medline on retrieval of human studies for knowledge synthesis

Nicole Askin; Tyler Ostapyk; Carla Epp

See end of article for authors' affiliations.

Objective: Use of the search filter 'exp animals/ not humans.sh' is a well-established method in evidence synthesis to exclude non-human studies. However, the shift to automated indexing of Medline records has raised concerns about the use of subject-heading-based search techniques. We sought to determine how often this string inappropriately excludes human studies among automated as compared to manually indexed records in Ovid Medline.

Methods: We searched Ovid Medline for studies published in 2021 and 2022 using the Cochrane Highly Sensitive Search Strategy for randomized trials. We identified all results excluded by the non-human-studies filter. Records were divided into sets based on indexing method: automated, curated, or manual. Each set was screened to identify human studies.

Results: Human studies were incorrectly excluded in all three conditions, but automated indexing inappropriately excluded human studies at nearly double the rate as manual indexing. In looking specifically at human clinical randomized controlled trials (RCTs), the rate of inappropriate exclusion of automated-indexing records was seven times that of manually-indexed records.

Conclusions: Given our findings, searchers are advised to carefully review the effect of the 'exp animals/ not humans.sh' search filter on their search results, pending improvements to the automated indexing process.

Keywords: Evidence Synthesis; Abstract and Indexing; Medical Subject Headings (MeSH); Automated Indexing

See end of article for supplemental content.

INTRODUCTION

Knowledge synthesis searching attempts to comprehensively retrieve all published literature on a particular question using a replicable search strategy. To address the volume of literature retrieved for broad or popular topics, information specialists often incorporate search filters to focus the search to particular types of records. Many of these filters designed for the Medline or PubMed databases use Medical Subject Heading (MeSH) controlled vocabulary terms. One particularly common strategy is the use of "double NOT" or exclusion filters to rapidly exclude irrelevant results based on subject headings. The Cochrane Highly Sensitive Search Strategies use the string 'exp animals/ not humans.sh' combined with NOT against the rest of the search strategy to limit searches to human studies [1]. Variants of this string have been widely adopted as standalone filters [2, 3] and as parts of other filters [4, 5].

The National Library of Medicine (NLM) implemented fully automated indexing of Medline in a graduated process, beginning with a pilot of 8 journals in 2019, 40% of journals in 2021, and 100% of journals by April 2022. Their method used the Medical Text Indexer-Automatic (MTIA), a natural language processing-based system, to assign MeSH terms [6]. MTIA identifies MeSH terms, synonyms, and trigger phrases in the title and abstract of records and incorporates position and frequency analysis in determining how to index an article [7]. In announcing the automated indexing transition, NLM highlighted improved timeliness and ability to scale indexing to meet the expanding volume of published literature as key drivers underlying the initiative [8]. However, anecdotal reports circulated among medical librarians and information specialists about failings in MeSH terms applied by the automated method, exemplified by the social media hashtag #meshfail. This raised questions

DOI: dx.doi.org/10.5195/jmla.2025.1972

about the continued reliability of MeSH-based searches and search filters.

While there is limited extant literature on the impact of automated indexing in Medline on information specialist practice, what does exist seems to bear out these concerns. For example, Hickner [9] conducted interviews with systematic searchers regarding search systems in which a respondent noted a concern about the impact of automated indexing on search precision. Chen and colleagues reported frequently missing or misused check tags, along with an apparent gender bias in ranking the Male/ heading over Female/ [10]. Koning and colleagues applied the MTI algorithm to texts from patent applications, and found that for the application of the Female/ subject heading the algorithm had a precision of 93% but recall of only 65% [11]. Most significantly, the work of Amar-Zifkin and colleagues identified multiple concerns with automated indexing in Medline, including irrelevant terms being included, obviously relevant terms being absent, and cases where better terms were available but unused [12, 13]. Causes they noted for these issues included misinterpretation of acronyms, rhetorical or metaphorical language, words that occur in multiple MeSH terms, and "unusual combinations of populations-interventions." They reported that nearly half of the records they examined exhibited one or more inadequacies.

In this study we sought to understand the impact of the switch to automated indexing on the use of purely MeSH-based filtering for knowledge synthesis. Specifically, we examined the use of the common exclusion filter 'exp animals/ not humans.sh', part of the Cochrane Highly Sensitive Search Strategies for identifying randomized trials [1]. We separated studies by indexing method—whether automated, curated or manual--and compared the frequency with which human studies were incorrectly excluded by this filter.

METHODS

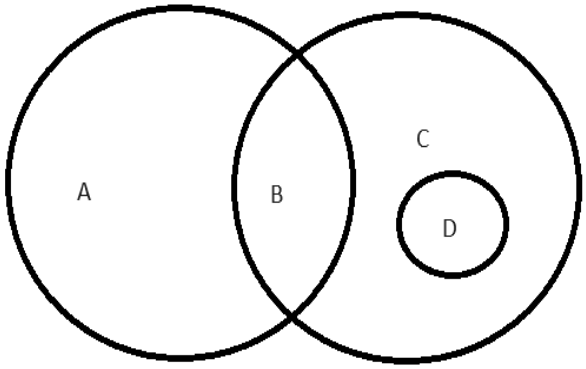
We searched Ovid Medline on 10 March 2023 using the Cochrane Highly Sensitive Search Strategy for identifying randomized trials in Medline, sensitivity- and precision-maximizing version, 2008 revision, Ovid format [1]. Results specifically excluded by the filter 'exp animals/ not humans.sh' were isolated and limited to Medline records with a publication year of 2021 and 2022 to capture a sample of records spanning the transition to fully automated indexing. Figure 1 shows the logic model for the Boolean used.

Figure 1 Venn diagram of Boolean logic model.

A+B are the studies which the filter is designed to retain: those tagged with humans/ in A, and those tagged with both humans/ and exp animals/ in B.

C represents the studies isolated for our analysis: studies tagged with exp animals/ but not humans/.

D represents the studies we sought to identify in that analysis: those studies within C that were human studies.



These results were then divided into sets based on the indexing method field (.ig). According to the Ovid database guide [14], there are three values that can be determined using this field:

- "Automated" reflects a record for which MeSH indexing is provided algorithmically
- "Curated" indicates that "MeSH indexing is provided algorithmically and a human reviewed (and possibly modified) the algorithm results"
- "Manual" is the designation we used for cases in which the field was not present, meaning "the indexing method is fully human indexed" [14].

This approach produced 4865 results for the automated set, 3062 for the curated set, and 2517 for the manual set. The full search strategy is available in Appendix A.

Each set was uploaded to a separate review in the knowledge synthesis platform Covidence, which automatically removed duplicates. The results were then screened blindly by two reviewers per record at both the title-abstract and full-text stages.

Records were included if they described a human study; no other study-type restrictions were imposed at the screening stage. For the purposes of this project, a very broad definition of human study was used, encompassing in vitro and ex vivo studies. Studies involving both humans and animals were considered human studies if one of the following cases applied: (1) the study was

DOI: dx.doi.org/10.5195/jmla.2025.1972

interventional and the intervention was performed on a human or human product (human cells, tissues, etc.); (2) the study was comparative and involved a comparison directly between groups of humans; or (3) in any study type, a significant outcome of interest was focused on humans. Studies involving human products administered to animal subjects were excluded unless the human product was cellular and was manipulated or analyzed in some way prior to administration. If a single paper described multiple studies, it was included if any of those studies met the definition of human study in use. For full-text screening of non-English studies, the Deepl translator was used to facilitate evaluation. The following exclusion rationales were applied in full-text screening:

- Animal study: the study is entirely on animals or animal products (animal cells, tissues, milk, etc.) and involves no human subjects or human products
- Wrong study type: the study does not have either human or animal subjects
- Ag/vet: the study is agricultural or veterinary in nature and does not meet any of the inclusion criteria to be considered human
- Human product: the study involved administration of an unmanipulated/unanalyzed human product to animals
- Unable to obtain full text through library subscriptions or interlibrary loan

PRISMA flow diagrams [15] for all three sets are included in Appendix B.

We created and pre-tested a custom data extraction form for analysis of the included studies. In designing the data extraction form we sought to include factors that may impact accuracy of indexing, including the presence of an abstract, the language of the full study, and whether the paper described multiple studies. We provided a list of study designs and study contexts (clinical, preclinical, educational, agricultural, veterinary, and other) with definitions to improve selection accuracy. We also extracted data related to the population, intervention, and indexing characteristics of the records. Finally, the form included an optional free-text field in which the extractor could propose potential reasons for the study to have been misinterpreted as an animal study. We used a two-reviewer model for the extraction process, with one reviewer serving as data extractor and the second as data verifier.

RESULTS

Each set of results included human studies based on the criteria set out in this project. We identified 205 human studies out of 4865 in the automated indexing set, 69 human studies out of 3062 in the curated indexing set, and 56 human studies out of 2517 in the manual indexing set. Thus, 4.2% of the articles in the automated indexing set were found to be human studies, compared to 2.3% in the curated set and 2.2% in the manual set.

We considered that the language of publication of the full-text study might impact the accuracy of the automated indexing compared to manual indexing, which includes access to full text. However, the manual set contained no non-English studies, so we were unable to assess the impact of language on this indexing method. The automated set had 522 non-English studies, 15 of which (2.9%) were classified as human studies, and the curated set had 238 non-English studies, 3 of which (1.3%) were classified as human studies.

Many of the human studies records described multiple studies: 53 out of 205 (25.9%) in the automated set, 35 out of 69 (50.7%) in the curated set, and 23 out of 56 (41.1%) in the manual set. For example, a single record might describe an animal RCT as well as a human epidemiologic study. Records in this category may have been appropriately indexed as animal studies but missed the addition of the humans/check tag. Similarly, there were some review articles which included analysis of both human and animal primary literature that were incorrectly indexed as solely animal studies. We conducted an analysis of the study type(s) considered a human study according to our definition. The distribution is shown in Table 1, although some multi-study records describe multiple human studies.

Table 1 Human studies by study type per set.

Type	Automated	Curated	Manual
RCT	85	14	17
NRE	44	29	16
Epidemiologic	20	12	3
Review	43	9	10
Qualitative	2	2	3
Case series/report	2	0	0
DTA	7	1	1
Economic	2	0	2
Opinion	1	0	0
Other	7	3	5

We additionally categorized included records according to study context: clinical, preclinical (laboratory), educational, and other. Because of our expansive definition of human studies, some records with an

DOI: dx.doi.org/10.5195/jmla.2025.1972

agricultural or veterinary context were also eligible for inclusion. As with study type, some records described multiple study contexts, although this was less frequently an issue. The distribution is presented in Table 2.

Table 2 Distribution of human studies by study context per set.

Context	Automated	Curated	Manual
Clinical	134	27	17
Preclinical	102	37	32
Educational	4	1	0
Agricultural	9	3	4
Veterinary	6	1	3
Other	4	2	1

Combining these two datasets allows us to determine the incorrect exclusion of human studies by type and context for each indexing method. We conducted a subanalysis on the clinical subset since this is the context in which the Cochrane RCT filter was developed. The exclusions by study type in the clinical context are displayed in Table 3.

Table 3 Exclusions of clinical-context records by study type and indexing method. Percentage of total set represents the entire set of articles excluded by the filter 'exp animals/ not humans.sh' for that indexing method.

	Automated		Curated		Manual	
	Count	% of total set (n=4865)	Count	% of total set (n=3062)	Count	% of total set (n=2517)
RCT	67	1.38%	7	0.23%	5	0.20%
NRE	7	0.14%	4	0.13%	0	0.00%
Epidemiologic	14	0.29%	7	0.23%	0	0.00%
Review	40	0.82%	8	0.26%	10	0.40%
Qualitative	1	0.02%	0	0.00%	0	0.00%
Case	1	0.02%	0	0.00%	0	0.00%
DTA	4	0.08%	1	0.03%	0	0.00%
Economic	0	0.00%	0	0.00%	2	0.08%
Opinion	1	0.02%	0	0.00%	0	0.00%
Other	1	0.02%	1	0.03%	1	0.04%

In addition to the quantitative analysis, we conducted a qualitative review of the record abstracts to see whether we could discern the reasons for which records may have been incorrectly indexed. The reasons are summarized in Table 4.

Table 4 Potential reasons for animal indexing, by indexing method.

	Automated		Curated		Manual	
	Count	% of included records (n=205)	Count	% of included records (n=69)	Count	% of included records (n=56)
Unknown	3	1.46%	1	1.45%	0	0.00%
Agricultural or veterinary context	16	7.80%	5	7.25%	8	11.59%
Allergy-related study	3	1.46%	0	0.00%	0	0.00%
Animal product (meat, milk, etc.)	39	19.02%	4	5.80%	2	2.90%
Animal model (e.g. for evaluating a surgical procedure)	5	2.44%	3	4.35%	3	4.35%
Human product administered to an animal population	9	4.39%	3	4.35%	8	11.59%
Mentions excluding animals	3	1.46%	0	0.00%	0	0.00%
Includes animal study	87	42.44%	43	62.32%	28	40.58%
Animal-related language (e.g. "click a mouse")	5	2.44%	2	2.90%	2	2.90%
Mentions prior animal work	32	15.61%	4	5.80%	0	0.00%
Animal-borne/zoonotic diseases	5	2.44%	4	5.80%	8	11.59%
Pet-related study	5	2.44%	0	0.00%	0	0.00%
Other	3	1.46%	2	2.90%	0	0.00%

DOI: dx.doi.org/10.5195/jmla.2025.1972

The most common reason for inappropriate indexing across all sets was that the record included animal studies – whether because the record described multiple studies, or for example, a review or opinion paper discussed both human and animal work. The comparatively high proportion of agricultural, veterinary, animal-disease, and human-product studies in the manually curated set is likely related to our particular definition of human studies.

However, some of the reasons for inappropriate indexing specific to the automated indexing set warrant further evaluation. Of the 39 studies that may have been excluded due to inclusion of an animal product, 17 (44%) involved a dietary intervention, suggesting that this topic of research may be significantly impacted by inappropriate indexing. Other animal-product exclusions concerned use of animal tissue in transplantation or in the development of vaccines.

DISCUSSION

Our analysis demonstrated that human clinical RCTs are excluded by the 'exp animals/ not humans.sh' filter in automated-indexing records at six times the rate of curated-indexing records and nearly seven times the rate of manually-indexed records. Concerningly, the mention of prior animal work in the abstract was a very common reason for inappropriate indexing, as the algorithm is unable to understand that this mention is not what the article is "about". Along those same lines, some records were indexed as animal studies when their abstracts specifically mentioned excluding animals (particularly in reviews). Finally, studies which included an animal-related intervention (such as a pet) or problem (such as an allergy) were indexed using terms related to the animal(s) involved but had no corresponding humans/ check tag.

Given our findings, we urge information specialists conducting knowledge synthesis projects in Medline or PubMed to exercise caution in using pure-MeSH filters, particularly the common filter 'exp animals/ not humans.sh'. In practical terms, the findings for human clinical RCTs suggest that this filter could remove one human study for every 100 automated-indexed records included in search results. This unintended removal of relevant evidence for knowledge synthesis projects is concerning. If possible, given result volumes and screening resources, we recommend against use of this filter until the automated indexing algorithm is improved. In particular, the filter should not be used for bodies of literature that are likely to use animal-related terminology – for example, studies of diets and dietary interventions.

The NLM has developed a new automated indexing system, termed MTIX, which is based on machine learning [6]. This system is asserted to significantly outperform MTIA in terms of accuracy: NLM testing found the F1 score (a combination assessment of recall and precision) for the human and animal check tags applied using MTIX to be 96% and 92% respectively [7]. In particular, NLM asserts that MTIX will be able to appropriately assess metaphorical language [7]. Future work will be necessary once this system has been fully implemented to assess whether it adequately addresses the problems of the MTIA system.

LIMITATIONS

This study was conducted solely in Medline and assesses the impact of automated indexing specifically in the context of MeSH. As automated application of subject heading terms is extended to other databases [16], testing in those contexts will be required to assess whether those databases experience a similar rate of indexing concerns. Additionally, although automated indexing using MTIA is based solely on title-abstract and therefore a lack of abstract could significantly affect indexing [7], it was not possible to assess the impact of a lack of abstract on the results since the search sets had only a total of 16 studies without abstracts between them.

In evaluating whether particular records described a human study, we used a very broad definition of the term, including in vitro and ex vivo studies. A narrower definition may have impacted our findings. We also found that in some cases it was quite challenging to conclusively categorize studies as human or not, and we recognize that there may be some variation in analysis, particularly in two areas: use of human products on animal subjects, and agricultural/educational studies. We adopted a double-screen method for both title-abstract and full-text screening in an effort to provide a consensus-based confirmation of categorization.

While our study identifies human studies across several study types, the sets used for this analysis were developed using a search strategy specifically designed to retrieve randomized controlled trials. As such, further testing is needed to verify the impact on human-study filtering in other study types or contexts. Similarly, we identified some specific topics where our results suggest inappropriate exclusion of results may be more common, but the impact is likely to be variable across other search topics. Finally, our study was specific to use of the animals/humans check tags, but we believe there would likely be similar issues with other pure MeSH-based filtering approaches; additional research is warranted to confirm this suspicion and evaluate the extent of the problem.

As we conducted data extraction, we also noted that after our initial search, fourteen records had been changed from an indexing method of automated to curated, often because a human subject heading was added. While we view this as positive evidence of improvement in indexing by NLM, for the purposes of analysis we retained such studies in their original set. As noted by Amar-Zifkin and

DOI: dx.doi.org/10.5195/jmla.2025.1972

colleagues, these types of changes to indexing have a negative, though slight, impact on replicability [13]. As such, while retroactive reindexing could correct the errors of MTIA, it could also in itself cause problems for knowledge syntheses.

STATEMENTS

The authors report no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data associated with this article are available via the University of Manitoba Dataverse: https://borealisdata.ca/dataverse/AutomatedIndexing.

AUTHOR CONTRIBUTIONS

Nicole Askin: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – Original draft, Writing – Review & editing, Supervision, Project administration. Tyler Ostapyk: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – Original draft, Writing – Review & editing, Supervision, Project administration. Carla Epp: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – Original draft, Writing – Review & editing, Supervision, Project administration.

REFERENCES

1. Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf MI, Noel-Storr A, Rader T, Shokraneh F, Thomas J, Wieland LS. Technical Supplement to Chapter 4: Searching for and selecting studies. In: Cochrane Handbook for Systematic Reviews of Interventions [Internet]. 6th ed. Cochrane; 2019. Available from: https://training.cochrane.org/handbook/version-6/chapter-4-tech-suppl.
2. CADTH. Animal studies - standard - MEDLINE, Embase [Internet]. Available from: https://searchfilters.cadth.ca/link/6.
3. Schulich Librarians. Search filters and tools [Internet]. McGill Library. [cited 2024 Mar 27]. Available from: https://libraryguides.mcgill.ca/knowledge-syntheses/search-tools.
4. Avau B, Van Remoortel H, De Buck E. Translation and validation of PubMed and Embase search filters for identification of systematic reviews, intervention studies, and observational studies in the field of first aid. J Med Libr Assoc. 109(4):599–608.
5. Search strategies [Internet]. BMJ Best Practice. [cited 2024 Mar 27]. Available from: https://bestpractice.bmj.com/info/toolkit/learn-ebm/study-design-search-filters/.
6. Krithara A, Mork JG, Nentidis A, Paliouras G. The road from manual to automatic semantic indexing of biomedical literature: a 10 years journey. Front Res Metr Anal. 2023 Sep 29;8:1250930.
7. NLM. NLM Office Hours: PubMed [Internet]. U.S. National Library of Medicine; [cited 2024 Mar 25]. Available from: https://www.nlm.nih.gov/oet/ed/pubmed/02-24_medline-automated-indexing.html.
8. NLM. MEDLINE 2022 Initiative: transition to automated indexing [Internet]. U.S. National Library of Medicine; [cited 2023 Nov 28]. Available from: https://www.nlm.nih.gov/pubs/techbull/tb.html.
9. Hickner A. How do search systems impact systematic searching? A qualitative study. J Med Libr Assoc. 111(4):774–82.
10. Chen E, Bullard J, Giustini D. Automated indexing using NLM's Medical Text Indexer (MTI) compared to human indexing in Medline: a pilot study. J Med Libr Assoc. 111(3):684–94.
11. Koning R, Samila S, Ferguson JP. Who do we invent for? Patents by women focus more on women's health, but few women get to invent. Science. 2021 Jun 18;372(6548):1345–8.
12. Amar-Zifkin A, Paquet V, Ekmekjian T, Landry T. Exploring the impact of automated indexing on completeness of MeSH terms. In: 2023 Canadian Health Libraries Association [Internet]. 2023 [cited 2023 Nov 28]. Available from: https://papyrus.bib.umontreal.ca/xmlui/handle/1866/28262.
13. Amar-Zifkin A, Ekmekjian T, Paquet V, Landry T. Brief communication – concerning algorithmic indexing in MEDLINE. J EAHIL. 2024 Mar 17;20(1):18–21.
14. Wolters Kluwer Heath. Medline 2024 database guide. Wolters Kluwer; [cited 2024 Jul 2]. Available from: https://ospguides.ovid.com/OSPguides/medline.htm
15. Page M, McKenzie J, Bossuyt P, Boutron I, Hoffmann T, Mulrow C, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. PLOS Med. 2021 Mar 29;18(3):e1003583.
16. Golub K. Automated subject indexing: an overview. Cat Classif Q. 2021 Nov 29;59(8):702-19.

SUPPLEMENTAL FILES

- **Appendix A**: Medline search strategy
- **Appendix B**: PRISMA flow diagrams

AUTHORS' AFFILIATIONS

Nicole Askin, Nicole.askin@umanitoba.ca, University of Manitoba, Manitoba, Canada

Tyler Ostapyk, tyler.ostapyk@umanitoba.ca, University of Manitoba, Manitoba, Canada

Carla Epp, Carla.epp@umanitoba.ca, University of Manitoba, Manitoba, Canada

Received April 2024; accepted July 2024

DOI: dx.doi.org/10.5195/jmla.2025.1972

ISSN 1558-9439 (Online)

ORIGINAL INVESTIGATION
DOI: dx.doi.org/10.5195/jmla.2025.1985

Evaluating a large language model's ability to answer clinicians' requests for evidence summaries

Mallory N. Blasingame; Taneya Y. Koonce; Annette M. Williams; Dario A. Giuse; Jing Su; Poppy A. Krump; Nunzia Bettinsoli Giuse, FMLA

See end of article for authors' affiliations.

Objective: This study investigated the performance of a generative artificial intelligence (AI) tool using GPT-4 in answering clinical questions in comparison with medical librarians' gold-standard evidence syntheses.

Methods: Questions were extracted from an in-house database of clinical evidence requests previously answered by medical librarians. Questions with multiple parts were subdivided into individual topics. A standardized prompt was developed using the COSTAR framework. Librarians submitted each question into aiChat, an internally managed chat tool using GPT-4, and recorded the responses. The summaries generated by aiChat were evaluated on whether they contained the critical elements used in the established gold-standard summary of the librarian. A subset of questions was randomly selected for verification of references provided by aiChat.

Results: Of the 216 evaluated questions, aiChat's response was assessed as "correct" for 180 (83.3%) questions, "partially correct" for 35 (16.2%) questions, and "incorrect" for 1 (0.5%) question. No significant differences were observed in question ratings by question category (p=0.73). For a subset of 30% (n=66) of questions, 162 references were provided in the aiChat summaries, and 60 (37%) were confirmed as nonfabricated.

Conclusions: Overall, the performance of a generative AI tool was promising. However, many included references could not be independently verified, and attempts were not made to assess whether any additional concepts introduced by aiChat were factually accurate. Thus, we envision this being the first of a series of investigations designed to further our understanding of how current and future versions of generative AI can be used and integrated into medical librarians' workflow.

Keywords: Large Language Models; LLMs; Generative AI; Artificial Intelligence; Evidence Synthesis; Library Science; Information Science; Biomedical Informatics

INTRODUCTION

Following the public launch of OpenAI's Chat Generative Pre-Trained Transformer (ChatGPT) in November 2022 [1], much consideration has been given in the academic and popular discourse to the current and anticipated impact of generative artificial intelligence (AI) on a number of professions. Within the health sciences, studies have investigated the ability of generative AI chat tools (including ChatGPT, Google Gemini, and Microsoft Copilot) to respond to patients' medical inquiries [2, 3], answer questions on licensing exams [4], support healthcare education [5], facilitate communication of research studies to lay audiences [6], aid with clinical documentation [7], and contribute to academic manuscripts [8], with many studies focused on specific specialty areas [4]. Much has also been written about the potential for librarians to become expert AI knowledge workers, who could play a critical role in advising and instructing library users on how to best use and integrate these AI tools for their information needs [9-12]. Although none of these studies were conducted in a medical library, many discuss its potential usefulness in that setting [10-16]. Outside the field of medical librarianship, Qureshi et al. [17] and Wang et al. [18] have explored how the application of generative AI could be developed to aid with search strategies and systematic reviews. None of the studies so far seem to have investigated the performance of generative AI in the critical task of searching and synthesizing knowledge from the medical literature, particularly in comparison with medical librarians' expertise in this area.

At the Center for Knowledge Management at Vanderbilt University Medical Center (VUMC), our team of medical librarians has, for over twenty years, provided evidence syntheses of the biomedical literature to respond to clinicians' questions, many of which are complex (i.e., questions containing clusters of questions), arising from clinical encounters. These questions were gathered initially through rounding with clinical teams and, since 2004, via an evidence request message basket service

DOI: dx.doi.org/10.5195/jmla.2025.1985

linked within the electronic health record (EHR) to facilitate clinicians' ability to send requests at the time and place when they most need an answer [19-23]. A previous study found high levels of physician satisfaction with the evidence summaries provided by our team [24]. This service requires librarians to be highly trained and able to quickly search and filter the current available literature on the topic, extract the most salient information needed to answer the question, and prepare a concise but comprehensive narrative synthesis that is returned to the clinician to inform decision-making [25]. Given the ability of generative AI chat tools to quickly produce detailed, fully articulated summaries drawn from a large body of knowledge, evaluating their current performance in responding to clinical questions is critical to understanding how they may eventually be integrated into medical librarians' workflows.

Some studies assessing generative AI tools' ability to provide comprehensive and accurate responses to clinical questions have observed that they can produce accurate results [26-29], particularly for less complex requests [26], although variation in results has been observed among different specialties, tasks, and models investigated [4]. Significant limitations have also been observed, including introduction of both minor and major errors via hallucination or misinterpretation [26, 30–31], lack of up-to-date information [32], and limited domain-specific content knowledge [33]. However, with ongoing updates and refinement, it is anticipated that these tools will continue to improve, with advancements already observed, for example, in comparisons of GPT-3.5 to GPT-4 [26, 34].

Previous studies have evaluated generative AI chat bots' responses to clinical questions in comparison with a) published practice guidelines [27, 35–37], b) objective multiple-choice answers [4], and/or c) assessment by clinical experts' review [4, 26, 38–40]. Our study uses a different approach in that, to our knowledge, no studies have yet evaluated generative AI tools in the context of responding to actual clinical questions that arise from patient healthcare encounters and use medical librarians' evidence syntheses as a reference standard. This study builds upon previous research in knowledge acquisition [41–42] and continues our examination of how AI could aid or eventually transform medical librarians' work [43–45]. Given our institutional policies restricting the use of publicly available generative AI tools, we used aiChat [46], a VUMC-managed generative AI tool.

This study aimed to investigate the current ability of aiChat to answer individual clinical questions compared to expertly trained librarians when questions are formulated in a standardized manner. Specifically, the study investigated the following questions:

1. How accurate are aiChat's responses to clinical questions, as compared with medical librarians' gold-standard evidence syntheses?
2. Is aiChat's performance affected by question adjudication status (i.e., whether a third person was needed to resolve discordant ratings by two independent reviewers)?
3. Are there significant differences in aiChat's performance by question category?
4. What proportion of references included in aiChat responses can be verified to exist?

The main rationale for undertaking this study is to provide the field of biomedical librarianship with sufficient elements of investigation to promote interest and curiosity towards AI and its potential usefulness in our field.

METHODS

Our team of information specialists has, for many years, received clinical questions from providers via rounding and via an evidence request message basket service linked within the electronic health record (EHR), providing evidence syntheses as a response. The evidence syntheses are created leveraging the comprehensive collection of journals at Vanderbilt University Medical Center, through freely available biomedical literature and grey literature published online, and via document delivery when needed. A sample of actual clinical questions was used to compare the accuracy of a locally managed generative AI chat tool's responses with librarians' gold-standard responses. Although these questions were generated by clinicians in connection to specific patient cases, they do not, by design, include any identifiable patient information, and the study was determined to be exempt by the Vanderbilt University Medical Center Institutional Review Board (IRB 240714). As applicable, this study adhered to the JAMA Network Guidance for Reporting Use of AI in Research and Scholarly Publication [47].

Generative AI Tool

As submission of proprietary data to public-facing generative AI tools is restricted by our medical center policy, we used an organizationally approved, internally managed AI chat tool called aiChat to conduct the study [46]. aiChat is an in-house instance of OpenAI's GPT-4 large language model made available to users at our institution behind a secure firewall using Microsoft Azure's cloud computing services [48]. At the time of the study, aiChat was a Beta version with options to use either OpenAI's GPT-3.5 or GPT-4 models. Similar to the public version of ChatGPT, aiChat allows users to choose the GPT model of interest, submit one or more prompts, and receive a response in a user-friendly, conversational format. For this study, we chose to use the GPT-4 model

DOI: dx.doi.org/10.5195/jmla.2025.1985

due to the improvements in accuracy, advanced reasoning, and its more extensive training data set compared to GPT 3.5 [49-50]. The model was used as provided by aiChat; no additional dataset was used to train the algorithm.

Question Pool

The in-house database used to assign, document, and archive clinicians' questions and the corresponding evidence synthesis responses provided by our team was queried to retrieve all questions received since 2010 [20, 21, 23, 51]. To align with GPT-4's most recent knowledge cutoff date at the time of the study, we excluded questions received after April 2023. A group of information scientists then determined eligibility of each archived question. The question set was limited to those that addressed a clinician's information need during patient care; general education questions and patient education requests were excluded. Questions were also excluded if the evidence synthesis response provided by the librarian contained just a list of citations, provided only an annotated list of citations, or reported that no answer was found in the literature. Prior to being assigned the task of establishing eligibility, all librarians worked on the same sample set of questions to determine consistency. As the eligibility criteria were clearly defined and easy to understand, the reviewers were quickly able to resolve the few minor differences in interpretation.

For this initial study, we aimed to assess aiChat's performance when responding to one focused question at a time, with future analyses planned to assess aiChat's performance with entire complex, multi-faceted requests comprised of clusters of questions. Therefore, requests containing more than one distinct question (e.g., both diagnosis and treatment) were divided into individual questions by information scientists in alignment with the methods established by Giuse et al. [52]. Each was considered a separate question for the study. In some cases, questions were reworded for clarity or to remove irrelevant information from the requestor's original message (e.g., details about requested turnaround time).

Determining Critical Elements of the Responses

The librarians' original evidence summaries were used as the gold standard for comparison with aiChat's response summaries. To facilitate the comparison, pairs of medical librarians reviewed the original evidence synthesis response for each included question and came to consensus on the concepts that were most critical (i.e., the part of text directly answering the question at hand) and necessary to answer each question. For this process, librarians focused on the most pertinent, high-level conclusions, in recognition that there may be wide variation in wording and other elements within narrative summaries that nonetheless reach the same conclusions. These critical elements were copied from the original response and recorded in REDCap [53, 54] alongside the question, prior to submitting the question to aiChat.

Prompt Engineering and Submission

Consultation of the literature for current practices for effective prompt engineering revealed no widely accepted, authoritative guidelines. However, researchers have suggested approaches to improve the quality of generative AI's response, which were consistent with observations from initial testing by our team, such as giving the chat bot a clear role, establishing the context, and defining the expected output in terms of format and audience [55–57]. The COSTAR framework (Context, Objective, Style, Tone, Audience and Response) [58] was selected to guide prompt engineering for this study as it provides specific details to inform the GPT response, including the use of delimiters to specify the input's distinct components, and incorporates many of the principles recommended in the literature [58–61]. Using the framework, senior members of the team with expertise in librarianship, knowledge acquisition, medicine, and artificial intelligence devised a standardized prompt to submit with each clinical question (Figure 1).

Figure 1 Standardized prompt, in COSTAR format, used to submit each question to aiChat.

COSTAR Prompt

#CONTEXT#
I am a medical librarian at a major academic health sciences center. In my team, our members provide evidence-based filtered summaries of the biomedical literature for use in patient care.

#OBJECTIVE#
Your task is to provide a summary of evidence that answers a clinical question I will provide to you. This involves scanning both the published and grey literature. The aim is to create a narrative statement that answers the question. When possible, the narrative statement should comment on the strengths and weaknesses of the evidence. Only use information that was available prior to [date of packet].

#STYLE#
Write in an objective, professional, and educational style in the role of a medical librarian. Write the response in a style that is directed towards medical professionals interested in understanding the available evidence.

#TONE#
Maintain a balanced and objective tone throughout the summary.

#AUDIENCE#
The target audience is clinicians providing patient care. Assume a readership that has direct experience in taking care of patients.

#RESPONSE FORMAT#
Provide an easy-to-follow narrative summary in paragraph format.

#START ANALYSIS#
If you understand, ask me to enter the clinical question.

COSTAR=Context, Objective, Style, Tone, Audience and Response [58].

OpenAI's GPT-4's training data set includes a variety of Internet sources, including books, articles, and websites; specific details are generally not available [62-63]. aiChat was prompted to only use data from its training set published prior to the date of the original clinician request

DOI: dx.doi.org/10.5195/jmla.2025.1985

for evidence to avoid inclusion of information that would have not yet been published, and thus not available to the librarian compiling the evidence summary. In testing, aiChat was able to adjust the response by date when given this parameter. Given that studies have established that GPT often fabricates references [14, 30, 64], the team did not specifically ask aiChat to provide references as part of the prompt. Providing an example of the desired output within the prompt has also been suggested [61] and found to improve performance in some analyses [18]. However, it is unlikely that a user asking a real clinical question would have an example response readily available to submit, so examples were not included in our prompt.

All questions were submitted to aiChat between March 25, 2024 – April 1, 2024. To capture aiChat's responses to each question, medical librarians worked in pairs to submit assigned sets of questions to the chat bot tool. First, a librarian selected "New prompt," set aiChat to use GPT-4, and submitted the prompt (Figure 1). When aiChat responded to confirm understanding (e.g., "Understood. Please enter the clinical question."), the clinical question was copied directly from the REDCap database and submitted within the same encounter. The full response from aiChat was copied from the interface and saved in REDCap.

Initially, a set of five randomly selected test questions was submitted to aiChat five times each back-to-back in sequence by a senior member of our team to assess whether there was enough variation in the responses to necessitate submitting each question multiple times. Although variance was observed in the wording and other elements of aiChat's summary replies, the overall concepts and conclusions were consistent. Other research has observed significant differences in ChatGPT's responses when prompts are submitted multiple times [18]. However, this study aimed to assess the performance of generative AI in the real-life scenario of a clinician seeking a response to a clinical question. In this context, submitting a question multiple times would not be practical. Thus, for this study, the team decided to submit each question only one time.

GPT Response Evaluation

Each question, along with the critical elements from the original packet and the response from aiChat, was assigned to two independent medical librarian reviewers to evaluate the extent to which aiChat's response aligned with the original librarian's gold-standard synthesis of evidence from the published and grey literature. Each reviewer independently assessed whether aiChat answered the question correctly in comparison with the original gold-standard response. The assessment was based on whether aiChat included all, some, or none of the key critical elements that were identified by consensus from the librarian's original summary. Reviewers used a 3-point Likert scale adapted from Suárez et al. [40] to indicate whether aiChat's overall response was incorrect (1), partially correct (2), or correct (3). See Table 1 for detailed descriptions of each grading. The response options avoid the use of non-numerical, vague qualitative terminology (e.g., "mostly correct"), as these types of phrases may create ambiguity and difficulty with interpretation [65–66]. For example, different reviewers may interpret and apply the concept of "mostly" in different ways. In our study, the determination of partially correct (2) was quantitatively determined by counting the number of absent critical elements (Table 1). To be considered correct, it was not necessary for aiChat to use the exact same language from the original summary, but rather for the response to be conceptually similar. In cases where aiChat provided information beyond what the librarian included, the responses were not considered incorrect. If the aiChat summary included facts not present in the librarian's summary, no efforts were made to assess their accuracy.

Table 1 3-point Likert scale score criteria for GPT response evaluation.

Reviewer's grading	Description
Incorrect (1)	The answer does not address any critical elements identified in the librarian's original evidence summary.
Partially Correct (2)	The answer addresses one or more, but not all, critical elements identified by medical librarians.
Correct (3)	The answer addresses all critical elements identified in the original evidence summary.

Adapted from Suárez et al. [40]

For example, one question included in the study asked for recommendations on hyponatremia treatment with free water restrictions and sodium chloride (NaCl) tablets. The critical elements identified from the librarian's gold-standard response were: 1. Guidelines suggested the use of oral NaCl combined with loop diuretics for syndrome of inappropriate antidiuretic hormone secretion (SIADH) patients with hyponatremia; 2. Guidelines also recommended fluid restriction and hypertonic saline use for hyponatremia; 3. Some recommendations also addressed fluid restriction for SIADH and other conditions without mentioning the use of NaCl tablets. Overall, the aiChat answer addressed all key elements and was categorized as correct (3). If one or two of the three critical elements had been absent, this answer would have

DOI: dx.doi.org/10.5195/jmla.2025.1985

been classified as partially correct (2). The absence of all critical elements would make the answer incorrect (1).

Adjudication of Discordant Ratings

Discordant ratings of aiChat's performance by the two independent librarian reviewers were resolved (adjudicated) by a third reviewer with medical knowledge and expertise in evidence synthesis, librarianship, knowledge acquisition, and extensive experience with adjudication in knowledge acquisition research [67–68]. The adjudicator thoroughly reviewed each question, the full original summary, the complete aiChat summary, and, if needed, the original supporting references. When relevant, association websites referred to by the aiChat tool were also consulted. Specifically, we consulted websites as they existed at the time the original question was asked by using the Internet Archive's Wayback Machine [69]. This process allowed us to confirm whether more recent knowledge that would not have been available to the librarian at the time of the original request may have been incorporated into aiChat's response and thus created discrepancies.

Question Categories

To allow comparison of performance by question type, each question was assigned by a medical librarian to one of eight distinct categories: Disease Etiology, Diagnostic Procedure, Differential Diagnosis, Disease Description, Disease Complication, Disease Prevention, Disease Prognosis, or Treatment. These categories were adapted from a previous study conducted by NBG [52, 70]. In cases of multi-faceted questions, each of the individual questions was assigned a category.

Reference Verification

Although the prompt did not specifically request the inclusion of references, many of aiChat's responses did contain academic references with combinations of author name, journal, and/or publication year. The assessment of accuracy was based only on aiChat's summary. A separate exploratory analysis was performed using a sub-sample of questions to verify if the references provided by aiChat were real or hallucinated.

For this analysis, a smaller sample of questions with responses that included citations was identified through random selection using a random number generator [71]; each question was assigned to a pair of librarians. The librarians reviewed the responses from aiChat and independently attempted to locate all cited references using the details provided (e.g., author name, article title, publication date). Our team searched PubMed, Google, Google Scholar, and journal websites to verify whether the references supplied by aiChat could be matched to a published source. We documented whether the citation was found or not found, and, if located, assessed its open access status. A single librarian checked whether the located references were cited in the original evidence summary packet.

Statistical Analysis

The ratings for all questions were stored in REDCap and analyzed descriptively using medians, ranges, and frequency. For each question, the absolute (n) and relative frequency (%) of ratings of incorrect (1), partially correct (2), and correct (3) were tabulated. For group comparisons of the categorical data, we used Wilcoxon Rank Sum, Kruskal-Wallis, or Fisher's Exact tests. The Wilcoxon Rank Sum test was used for the analysis of nonparametric ordinal ratings when compared across two independent groups. The Kruskal-Wallis test was used for analysis of nonparametric ordinal ratings when compared across more than two independent groups, and Fisher's Exact test was used to evaluate nominal ratings across more than two independent groups All analyses were conducted with GraphPad Prism 10 software. A two-tailed p-value <0.05 was used as the threshold for statistical significance.

RESULTS

The study included 217 discrete questions. During adjudication, one question was excluded due to misclassification as a patient care-related question. The final number of questions analyzed for the study was 216.

Evaluation of the Accuracy of aiChat's Responses

Table 2 shows the overall ratings of the accuracy of the tool's responses compared with the medical librarians' gold-standard evidence syntheses. Overall, 180 (83.3%) of aiChat responses were assessed as correct in comparison with the original librarian's response, while 35 (16.2%) were assessed as partially correct and 1 (0.5%) was assessed as incorrect.

Performance by Adjudication Status

Consensus was achieved between librarian pairs on 182 (84.3%) of the responses; the remaining 34 (15.7%) responses required adjudication. Results were similar for responses requiring and not requiring adjudication, with 84.1% (n=153) of questions without adjudication and 79.4% (n=27) of questions with adjudication assessed as correct. The Wilcoxon Rank Sum test revealed there were no statistically significant differences in the ratings of responses that received adjudication in comparison to those that did not undergo adjudication (p=0.61). Of the adjudicated questions, most (n=32) were due to a discrepancy of one point (e.g., scores of partially correct [2] and correct [3]). Two questions were adjudicated due to a discrepancy between incorrect (1) and correct (3) scores.

DOI: dx.doi.org/10.5195/jmla.2025.1985

Comparison by Question Category

The most common question category was Treatment (n=147; 68.1%), which included topics such as treatment adverse effects and treatment efficacy, while the least commonly assigned category was Differential Diagnosis (n=1; 0.46%). The percent of aiChat responses assessed as correct was ≥80% across all categories. No significant differences were observed in the question ratings by category when evaluated by the Kruskal-Wallis test (p=0.73), nor were any patterns or trends identified. For a full report of results by each category, see Table 3.

Comparison by Adjudication and Question Category

The questions sent for adjudication at the highest proportion were related to disease prevention (n=2; 28.6%); none of the differential diagnosis questions were adjudicated (Table 4). No patterns were observed in the data, and there were no significant differences by category of questions that received adjudication when compared by Fisher's Exact test to questions that were not adjudicated (p=0.90).

Table 2 Ratings of aiChat's responses to discrete questions, by adjudication status.

Questions	Incorrect (1)	Partially Correct (2)	Correct (3)	Total
Questions without adjudication	1 (0.5%)	28 (15.4%)	153 (84.1%)	182 (84.3%)
Questions with adjudication	0 (0.0%)	7 (20.6%)	27 (79.4%)	34 (15.7%)
Total	**1 (0.5%)**	**35 (16.2%)**	**180 (83.3%)**	**216 (100%)**

Table 3 Ratings of aiChat's responses to discrete questions, by question category.

Question Category	Number of Questions	Incorrect (1)	Partially Correct (2)	Correct (3)
Disease Etiology	20	1 (5.0%)	1 (5.0%)	18 (90.0%)
Diagnostic Procedure	10	0 (0.0%)	2 (20.0%)	8 (80.0%)
Differential Diagnosis	1	0 (0.0%)	0 (0.0%)	1 (100%)
Disease Description	10	0 (0.0%)	1 (10.0%)	9 (90.0%)
Disease Complication	8	0 (0.0%)	0 (0.0%)	8 (100%)
Disease Prevention	7	0 (0.0%)	1 (14.3%)	6 (85.7%)
Disease Prognosis	13	0 (0.0%)	1 (7.7%)	12 (92.3%)
Treatment*	147	0 (0.0%)	29 (19.7%)	118 (80.3%)
Total	**216**	**1 (0.5%)**	**35 (16.2%)**	**180 (83.3%)**

**aggregates the treatment, treatment adverse effects, and treatment efficacy question categories*

Table 4 Adjudication status of aiChat's responses to discrete questions, by question category.

Question Category	Number of Questions	No Adjudication	Adjudication
Diagnosis Etiology	20	18 (90.0%)	2 (10.0%)
Diagnostic Procedure	10	8 (80.0%)	2 (20.0%)
Differential Diagnosis	1	1 (100%)	0 (0%)
Disease Description	10	9 (90.0%)	1 (10.0%)
Disease Complication	8	7 (87.5%)	1 (12.5%)
Disease Prevention	7	5 (71.4%)	2 (28.6%)
Disease Prognosis	13	12 (92.3%)	1 (7.7%)
Treatment*	147	122 (83.0%)	25 (17.0%)
Total	**216**	**182 (84.2%)**	**34 (15.8%)**

**aggregates the treatment, treatment adverse effects, and treatment efficacy question categories*

Verification of References from GPT Response

Though the prompt we used to submit the clinical questions to aiChat did not specifically ask citations to be included, the responses provided by the GPT often did include references. Sixty-six (30%) answers which included 162 references were randomly selected for citation verification. The number of references provided by aiChat per question ranged from 1-4, with a median of 2.45. Our team was able to verify the existence of 60 of the 162 references (37.0%). Most of the verifiable citations were indexed in PubMed (n=56; 93.3%), with the remaining available on the cited journal's website (n=2; 3.3%), a professional organization's website (n=1;1.67%) and the website of the Food and Drug Administration

DOI: dx.doi.org/10.5195/jmla.2025.1985

(n=1;1.67%). Of these 60 references, 35 were open access. Nineteen references (31.7%), all open access, overlapped with some of the citations used by the librarians in answering 14 questions.

DISCUSSION

In this initial study comparing generative AI summaries with medical librarians' gold-standard clinical evidence syntheses in response to individual clinical questions, an organizationally managed generative AI chat tool using GPT-4 was able to report key elements identified in the librarian's evidence synthesis for the majority of clinical questions examined. These results are promising but only a first step in what we foresee to be a series of many investigations into generative AI tools' ability to summarize the evidence to answer clinical questions. We recognize the complexity and responsibility of creating a valid, comprehensive, and trustworthy evidence synthesis and are cognizant of many of the issues discussed in an article from Zhang and colleagues, including the need to ensure that large language models are trustworthy, transparent, secure, and avoid perpetuating biases [72].

In our sample of clinical questions, aiChat provided a correct response for 83.3% of questions and a partially correct response for 16.2%, resulting in an overall 99.5% of questions having at least a partially correct response. Most of the questions in our study (68.1%) were treatment-related, which is consistent with the types of questions most frequently asked by clinicians [52, 70, 73]. No significant differences in accuracy were observed across different categories of clinical questions or adjudication status. The one summary rated by the reviewers as incorrect was a response to a question about genetic mutations associated with a particular disease, for which aiChat's response referenced a different gene than the one reported in the gold-standard evidence packet. This finding could possibly suggest a need to better understand how generative AI tools handle genetic information given the complexity of the field.

While the aiChat- and medical librarian-developed summaries were consistent overall in terms of the key concepts included, many (63%) of the supporting references included in a subsample of aiChat's responses could not be independently verified. The inability to trust references provided by large language models and, consequently, to be able to verify specific details and results of the studies cited in the responses they provide is currently a significant limitation to their use. However, it is possible that generative AI tools' performance in this area could improve as we continue to see a rise in open access publishing [72, 74–75] and the models are not as limited by subscription paywalls. Furthermore, the wider availability of open access resources may make it easier to fully trace and verify the sources underlying generative AI's responses [62–63]. Issues of copyright are also well-discussed in the literature; this remains a key issue, as researchers are largely unable to determine the full set of content used to train the large language models [32, 76–78]. Although it is important to note that in some instances, the information in the response could be drawn entirely from freely available abstracts, the lack of transparency on the details of the datasets still poses real concerns.

We also anticipate that GPT may improve its response if provided with a curated set of full-text articles selected by a medical librarian. The ability of AI tools to allow users to enter content could greatly improve the very controversial and troubling problem of reference hallucination [14, 30, 64]. Providing a new generation of generative AI tools with selected content may also aid in addressing ethical concerns when using large language models, which reflect the social biases and inequities present in the clinical research studies and other content included in their training sets [72, 79–80]. By selecting content to provide to the generative AI tool, we could additionally ensure that copyright issues are addressed [32, 76–78] and that content with curated references is fully representative of a diverse population and as free as possible from bias.

Tang et al. [31] conducted a study using ChatGPT and GPT-3.5 in which the generative AI tools were provided with content from Cochrane review abstracts from six clinical areas and prompted to provide four-sentence summaries of the systematic reviews. The study found that, in this context, the summaries included few instances of fabrication; however, errors (e.g., those related to misinterpretation of the content) were still observed. In November 2023, OpenAI introduced a feature allowing users to create custom GPTs through which they can provide their own knowledge (e.g., full-text articles or other written documents) for GPT to use when responding to prompts [81]. At the time of the study, this feature was not available through our organization's internal generative AI tool, but OpenAI does offer the ability to create custom GPTs at the Enterprise level to enable organizations to leverage this option with proprietary information. Tools harnessing generative AI to search and summarize academic papers using underlying literature databases (e.g., Consensus [82] and Scopus AI [83]) are also becoming available. Additional studies are needed in this area to fully understand current models' ability to accurately summarize research when provided with selected, full-text source material.

In addition to assessing generative AI tools' performance relative to that of humans, Shah and colleagues have also emphasized the importance of evaluating the benefits of large language models and considering how they can be leveraged to enhance our work rather than simply replicating it [84]. In this study, we observed that a strength of the aiChat responses was the formatting of the narrative summaries, which typically began with a brief introduction to the topic, followed by a well-organized

DOI: dx.doi.org/10.5195/jmla.2025.1985

summary with a balanced representation of the viewpoints found in the literature, and ended with brief conclusions. While the requestor receiving the evidence synthesis may be an expert who is already familiar with the topic, they may also wish to share the summary to educate other members of the team with varying specialties (e.g., pharmacists, nutritionists) or who may be more junior (e.g., medical students). Our team recognizes that the approach of establishing the background at the beginning of the response has educational value in our academic setting and considers the inclusion of all viewpoints in the literature to be a best practice for evidence synthesis [70, 85]. The organization used by aiChat to structure the responses also has educational value for our profession as a model that can be applied for instructional purposes to train clinical librarians.

A review article by Lund et al. [13] on how librarians in different fields and specialties could incorporate generative AI in their work also reports on interesting opportunities for use in medical librarianship, like using the AI tool as a digital assistant and research partner in a variety of information specialist roles. Mughari and colleagues [16] more specifically adds the potential for AI to be used by librarians in collection development, digital content curation, indexing, and data analytics, among other things. Sutton and Parisi [14], Roth and Wermer-Colan [15], and Friesen et al. [11] see generative AI as playing a role in searching and the systematic review process. Liu and colleagues [10], Zhang [9], Friesen et al. [11], and Epstein [12] additionally discuss how AI technologies represent a new opportunity for user instruction and training by librarians. It is undoubtedly the case that having librarians, in their role of educated knowledge workers, become active players in this AI revolution could provide them with a truly transformative way of incorporating this new technology in their profession.

LIMITATIONS

An assumption of this study was that medical librarians' original evidence syntheses accurately reflected the literature as of the original request date, and that clinicians who received the response trusted and agreed that the supporting evidence provided by the librarian answered their questions. Although we did not independently re-verify the information provided in these evidence syntheses, previous studies have found high levels of physician satisfaction with our team's evidence services [24].

Similarly, we did not assess the accuracy of every detail of aiChat's summary but rather focused on whether the most critical elements of the librarian's original response were present and, for a subset of questions, whether references could be verified to exist. No attempts were made in this study to evaluate whether any additional facts introduced by aiChat were accurate or whether the verified references were cited appropriately, as the comparison was based on whether the critical elements identified in the librarian's gold-standard response were included in aiChat's answer.

It is important to note, that this study was conducted at a single large academic medical center, and that may be reflected in the type/complexity of questions included in our dataset. Thus, the study may not be entirely generalizable to other environments.

For this study, we intentionally divided complex, multi-faceted questions (i.e., a question which includes a cluster of questions) into individual questions and separately evaluated aiChat's response to each question. aiChat's performance in responding to complex, multi-faceted questions taken as a whole was not evaluated.

Finally, it is possible that aiChat's performance was impacted by elements of prompt design, such as the lack of examples in the prompt or our decision to only submit each question once.

CONCLUSIONS

The findings of this study highlight the promising performance of a generative AI tool using GPT-4 for providing responses to individual clinical questions, while also confirming known limitations, such as reference fabrication. Our approach is replicable for other institutions who may wish to conduct similar investigations, but it would not apply to institutions where organizational policies do not allow the use of publicly available generative AI tools and do not provide their own internal versions. Since the aim of this study was to evaluate whether aiChat was able to answer clinical questions with a response which included the answer given by our established gold standard, we intentionally did not evaluate any additional conceptual differences in the summaries.

Additional avenues for future research include exploring generative AI's ability to respond to questions for which librarians found no answer and evaluating aiChat's answers to multi-faceted clinical questions. We also plan to conduct research 1) evaluating performance across multiple versions of GPT models to understand how they continue to evolve and improve over time, and 2) investigating which of the generative AI usage recommendations listed in our discussion above could most successfully be carried out in our environment. Given the current inability to independently verify many of the sources used for the generative AI responses, an important next step will be to conduct a more detailed analysis of the source material. A particular area of interest is to establish a better understanding of the extent to which questions can be answered through freely available open source literature. It will also be critical to understand how generative AI performance may improve

DOI: dx.doi.org/10.5195/jmla.2025.1985

when provided with a body of literature curated by expert medical librarians. This model could potentially couple GPT's strengths in summary generation with librarians' critical expertise in literature selection and assessment.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Spencer DesAutels and Sheila Kusnoor for their review and feedback on the manuscript.

FUNDING STATEMENT

Support for the REDCap database, used in this study for data entry and data collection, was provided by CTSA award UL1TR000445 from the National Center for Advancing Translational Sciences.

COMPETING INTEREST STATEMENT

The authors declare no competing interests for this study.

DATA AVAILABILITY STATEMENT

The clinical questions used in this study are not publicly available as the data is institutional proprietary information.

AUTHOR CONTRIBUTIONS

Mallory N. Blasingame: Methodology; investigation; visualization; writing—original draft; writing—review and editing. Taneya Y. Koonce: Methodology; investigation; data curation; formal analysis; visualization; writing—original draft; writing—review and editing. Annette M. Williams: Methodology; investigation; data curation; visualization; writing—review and editing. Dario A. Giuse: Methodology; investigation; writing—original draft; writing—review and editing. Jing Su: Methodology; investigation; writing—review and editing. Poppy A. Krump: Methodology; investigation; writing—review and editing. Nunzia Bettinsoli Giuse: Conceptualization; methodology; investigation; formal analysis; visualization; writing—original draft; writing—review and editing; supervision.

REFERENCES

1. OpenAI. Introducing ChatGPT [Internet]. OpenAI; 2022 Nov 30 [cited 2024 Apr 25]. <https://openai.com/blog/chatgpt>.
2. Johns WL, Kellish A, Farronato D, Ciccotti MG, Hammoud S. ChatGPT can offer satisfactory responses to common patient questions regarding elbow ulnar collateral ligament reconstruction. Arthrosc Sports Med Rehabil. 2024 Apr;6(2):100893. DOI: https://doi.org/10.1016/j.asmr.2024.100893
3. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023 Jun 1;183(6):589–96. DOI: https://doi.org/10.1001/jamainternmed.2023.1838
4. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. J Biomed Inform. 2024 Mar;151:104620. DOI: https://doi.org/10.1016/j.jbi.2024.104620
5. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, Househ M. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform. 2023 Jun 29;305:644–7. DOI: https://doi.org/10.3233/shti230580
6. Shyr C, Grout RW, Kennedy N, Akdas Y, Tischbein M, Milford J, Tan J, Quarles K, Edwards TL, Novak LL, White J, Wilkins CH, Harris PA. Leveraging artificial intelligence to summarize abstracts in lay language for increasing research accessibility and transparency. J Am Med Inform Assoc. 2024 Jul 15:ocae186. DOI: https://doi.org/10.1093/jamia/ocae186.
7. Dubinski D, Won SY, Trnovec S, Behmanesh B, Baumgarten P, Dinc N, Konczalla J, Chan A, Bernstock JD, Freiman TM, Gessler F. Leveraging artificial intelligence in neurosurgery-unveiling ChatGPT for neurosurgical discharge summaries and operative reports. Acta Neurochir (Wien). 2024 Jan 26;166(1):38. DOI: https://doi.org/10.1007/s00701-024-05908-3
8. Lechien JR, Gorton A, Robertson J, Vaira LA. Is ChatGPT-4 accurate in proofread a manuscript in otolaryngology-head and neck surgery? Otolaryngol--Head Neck Surg. 2023 Sep 17. DOI: https://doi.org/10.1002/ohn.526
9. Zhang B. ChatGPT, an opportunity to understand more about language models. Med Ref Serv Q. 2023 Apr-Jun;42(2):194-201. DOI: https://doi.org/10.1080/02763869.2023.2194149
10. Liu H, Azam M, Bin Naeem S, Faiola A. An overview of the capabilities of ChatGPT for medical writing and its implications for academic integrity. Health Info Libr J. 2023 Dec;40(4):440-446. DOI: https://doi.org/10.1111/hir.12509
11. Friesen E, Tanna H, Roy A. Artificial intelligence in subject-specific library work: trends, perspectives, and opportunities. Can J Acad Librariansh. 2023 Aug 1;9:1–26. DOI: https://doi.org/10.33137/cjal-rcbu.v9.39951
12. Epstein HB. Suggestions from experience and AI tools to teach evidence based practice to nurses. Med Ref Serv Q. 2024 Jan-Mar;43(1):59-71. DOI: https://doi.org/10.1080/02763869.2024.2289335
13. Lund BD, Khan D, Yuvaraj M. ChatGPT in medical libraries, possibilities and future directions: An integrative review. Health Info Libr J. 2024 Mar;41(1):4-15. DOI: https://doi.org/10.1111/hir.12518
14. Sutton A, Parisi V. ChatGPT: game-changer or wildcard for systematic searching? Health Info Libr J. 2024 Mar;41(1):1-3. DOI: https://doi.org/10.1111/hir.12517
15. Roth S, Wermer-Colan A. Machine learning methods for systematic reviews: a rapid scoping review. Dela J Public

DOI: dx.doi.org/10.5195/jmla.2025.1985

Health. 2023 Nov 30;9(4):40-47. https://pubmed.ncbi.nlm.nih.gov/38173960/.

16. Mughari S, Rafique GM, Ali MA. Effect of AI literacy on work performance among medical librarians in Pakistan. J Acad Librariansh. 2024 Sep 1;50(5):102918. DOI: https://doi.org/10.1016/j.acalib.2024.102918

17. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? Syst Rev. 2023 Apr 29;12(1):72. DOI: https://doi.org/10.1186/s13643-023-02243-z

18. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good Boolean query for systematic review literature search? [Internet]. arXiv; 2023 Feb 9 [cited 2024 Apr 25]. <http://arxiv.org/abs/2302.03495>.

19. Giuse NB, Kusnoor SV, Koonce TY, Ryland CR, Walden RR, Naylor HM, Williams AM, Jerome RN. Strategically aligning a mandala of competencies to advance a transformative vision. J Med Libr Assoc JMLA. 2013 Oct;101(4):261–7. DOI: https://doi.org/10.3163/1536-5050.101.4.007

20. Giuse NB, Koonce TY, Jerome RN, Cahall M, Sathe NA, Williams A. Evolution of a mature clinical informationist model. J Am Med Inform Assoc JAMIA. 2005;12(3):249–55. DOI: https://doi.org/10.1197/jamia.m1726

21. Giuse NB, Williams AM, Giuse DA. Integrating best evidence into patient care: a process facilitated by a seamless integration with informatics tools. J Med Libr Assoc JMLA. 2010 Jul;98(3):220–2. DOI: https://doi.org/10.3163/1536-5050.98.3.009

22. Blasingame MN, Williams AM, Su J, Naylor HM, Koonce TY, Epelbaum MI, Kusnoor SV, Fox ZE, Lee P, DesAutels SJ, Frakes ET, Giuse NB. Bench to bedside: detailing the catalytic roles of fully integrated information scientists. Presented at: Special Libraries Association Annual Conference, Cleveland, OH; June 18, 2019.

23. Koonce TY, Giuse DA, Blasingame MN, Su J, Williams AM, Biggerstaff PL, Osterman T, Giuse NB. Personalization of evidence: using intelligent datasets to inform the process. Presented at: Annual Meeting of the American Medical Informatics Association (virtual); November 2020.

24. Mulvaney SA, Bickman L, Giuse NB, Lambert EW, Sathe NA, Jerome RN. A randomized effectiveness trial of a Clinical Informatics Consult Service: impact on evidence-based decision-making and knowledge implementation. J Am Med Inform Assoc JAMIA. 2008;15(2):203–11. DOI: https://doi.org/10.1197/jamia.m2461

25. Giuse NB. Advancing the practice of clinical medical librarianship. Bull Med Libr Assoc. 1997 Oct;85(4):437–8.

26. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, Berkowitz ST, Finn AP, Jahangir E, Scoville EA, Reese TS, Friedman DL, Bastarache JA, van der Heijden YF, Wright JJ, Ye F, Carter N, Alexander MR, Choe JH, Chastain CA, Zic JA, Horst SN, Turker I, Agarwal R, Osmundson E, Idrees K, Kiernan CM, Padmanabhan C, Bailey CE, Schlegel CE, Chambless LB, Gibson MK, Osterman TJ, Wheless LE, Johnson DB. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023 Oct 2;6(10):e2336483. DOI: https://doi.org/10.1001/jamanetworkopen.2023.36483

27. Cakir H, Caglar U, Yildiz O, Meric A, Ayranci A, Ozgor F. Evaluating the performance of ChatGPT in answering questions related to urolithiasis. Int Urol Nephrol. 2024 Jan;56(1):17–21. DOI: https://doi.org/10.1007/s11255-023-03773-0

28. Ozgor BY, Simavi MA. Accuracy and reproducibility of ChatGPT's free version answers about endometriosis. Int J Gynaecol Obstet. 2024 May;162(2):691-695. DOI: https://doi.org/10.1002/ijgo.15309

29. Kuşcu O, Pamuk AE, Sütay Süslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? Front Oncol. 2023;13:1256459. DOI: https://doi.org/10.3389/fonc.2023.1256459

30. Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. Mayo Clin Proc Digit Health. 2023 Sep;1(3):226–34. DOI: https://doi.org/10.1016/j.mcpdig.2023.05.004

31. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, Xu Z, Ding Y, Durrett G, Rousseau JF, Weng C, Peng Y. Evaluating large language models on medical evidence summarization. Npj Digit Med. 2023 Aug 24;6(1):158. DOI: https://doi.org/10.1038/s41746-023-00896-7

32. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthc Basel Switz. 2023 Mar 19;11(6):887. DOI: https://doi.org/10.3390/healthcare11060887

33. Thapa S, Adhikari S. ChatGPT, Bard, and large language models for biomedical research: opportunities and pitfalls. Ann Biomed Eng. 2023 Dec;51(12):2647–51. DOI: https://doi.org/10.1007/s10439-023-03284-0

34. Liu J, Zheng J, Cai X, Wu D, Yin C. A descriptive study based on the comparison of ChatGPT and evidence-based neurosurgeons. iScience. 2023 Sep 15;26(9):107590. DOI: https://doi.org/10.1016/j.isci.2023.107590

35. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. J Med Internet Res. 2023 Dec 28;25:e51580. DOI: https://doi.org/10.2196/51580

36. Maksimoski M, Noble AR, Smith DF. Does ChatGPT answer otolaryngology questions accurately? The Laryngoscope. 2024 Mar 28. DOI: https://doi.org/10.1002/lary.31410

37. Blacker SN, Kang M, Chakraborty I, Chowdhury T, Williams J, Lewis C, Zimmer M, Wilson B, Lele AV. Utilizing artificial intelligence and chat generative pretrained transformer to answer questions about clinical scenarios in neuroanesthesiology. J Neurosurg Anesthesiol. 2023 Dec 19. DOI: https://doi.org/10.1097/ana.0000000000000949

38. Kerbage A, Kassab J, El Dahdah J, Burke CA, Achkar JP, Rouphael C. Accuracy of ChatGPT in common gastrointestinal diseases: impact for patients and providers. Clin Gastroenterol Hepatol. 2023 Nov 19;S1542-

DOI: dx.doi.org/10.5195/jmla.2025.1985

3565(23)00946-1. DOI: https://doi.org/10.1016/j.cgh.2023.11.008

39. Azadi A, Gorjinejad F, Mohammad-Rahimi H, Tabrizi R, Alam M, Golkar M. Evaluation of AI-generated responses by different artificial intelligence chatbots to the clinical decision-making case-based questions in oral and maxillofacial surgery. Oral Surg Oral Med Oral Pathol Oral Radiol. 2024 Mar 6;S2212-4403(24)00095-6. DOI: https://doi.org/10.1016/j.oooo.2024.02.018

40. Suárez A, Jiménez J, Llorente De Pedro M, Andreu-Vázquez C, Díaz-Flores García V, Gómez Sánchez M, Freire Y. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Comput Struct Biotechnol J. 2024 Dec;24:46–52. DOI: https://doi.org/10.1016/j.csbj.2023.11.058

41. Miller RA, Giuse NB. Medical knowledge bases. Acad Med J Assoc Am Med Coll. 1991 Jan;66(1):15–7. DOI: https://doi.org/10.1097/00001888-199101000-00004

42. Giuse DA, Giuse NB, Miller RA. A tool for the computer-assisted creation of QMR medical knowledge base disease profiles. Proc Symp Comput Appl Med Care. 1991;978–9.

43. Su J, Blasingame MN, Zhao J, Clark JD, Koonce TY, Giuse NB. Using a performance comparison to evaluate four distinct AI-assisted citation screening tools. Presented at: Medical Library Association Annual Meeting, Portland, OR; May 2024

44. Blasingame MN, Su J, Zhao J, Clark JD, Koonce TY, Giuse NB. Using a semi-automated approach to update clinical genomics evidence summaries. Presented at: Medical Library Association and Special Libraries Association Annual Meeting, Detroit, MI; May 18, 2023.

45. Koonce TY, Blasingame MN, Williams AM, Clark JD, DesAutels SJ, Giuse DA, Zhao J, Su J, Naylor HM, Giuse NB. Building a scalable knowledge management approach to support evidence provision for precision medicine. Presented at: AMIA Informatics Summit, Chicago, IL; March 2022.

46. Department of Biomedical Informatics. Generative AI at VUMC [Internet]. Vanderbilt University Medical Center; [cited 2024 Apr 25]. <https://www.vumc.org/dbmi/GenerativeAI>.

47. Flanagin A, Pirracchio R, Khera R, Berkwits M, Hswen Y, Bibbins-Domingo K. Reporting use of AI in research and scholarly publication-JAMA Network Guidance. JAMA. 2024 Apr 2;331(13):1096–8. DOI: https://doi.org/10.1001/jama.2024.3471

48. Miliard M. Vanderbilt piloting Nuance CAX Copilot, testing other genAI use cases [Internet]. Healthcare IT News; 2024 Jan 8 [cited 2024 Jul 16]. <https://www.healthcareitnews.com/news/vanderbilt-piloting-nuance-dax-copilot-testing-other-genai-use-cases>.

49. Terrasi V. GPT-4: how is it different from GPT-3.5? SearchEngineJournal; 2023 Mar 22 [cited 2024 Jul 16]. <https://www.searchenginejournal.com/gpt-4-vs-gpt-3-5/482463/>.

50. Griffith E. GPT-4 vs. ChatGPT-3.5: what's the difference? PCMag; 2023 Mar 16 [cited 2024 Jul 16]. <https://www.pcmag.com/news/the-new-chatgpt-what-you-get-with-gpt-4-vs-gpt-35>.

51. Giuse NB, Kafantaris SR, Miller MD, Wilder KS, Martin SL, Sathe NA, Campbell JD. Clinical medical librarianship: the Vanderbilt experience. Bull Med Libr Assoc. 1998 Jul;86(3):412–6.

52. Giuse NB, Huber JT, Giuse DA, Brown CW, Bankowitz RA, Hunt S. Information needs of health care professionals in an AIDS outpatient clinic as determined by chart review. J Am Med Inform Assoc JAMIA. 1994;1(5):395–403. DOI: https://doi.org/10.1136/jamia.1994.95153427

53. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 Apr;42(2):377–81. DOI: https://doi.org/10.1016/j.jbi.2008.08.010

54. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, McLeod L, Delacqua G, Delacqua F, Kirby J, Duda SN, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform. 2019 Jul;95:103208. DOI: https://doi.org/10.1016/j.jbi.2019.103208

55. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng. 2023 Dec;51(12):2629–33. DOI: https://doi.org/10.1007/s10439-023-03272-4

56. Ge J, Chen IY, Pletcher MJ, Lai JC. How I approach it: prompt engineering for generative artificial intelligence (GAI) in gastroenterology and hepatology. Am J Gastroenterol. 2024 Mar 20. DOI: https://doi.org/10.14309/ajg.0000000000002689

57. Lo LS. The art and science of prompt engineering: a new literacy in the information age. Internet Ref Serv Q. 2023 Oct 2;27(4):203–10. DOI: https://doi.org/10.1080/10875301.2023.2227621

58. Bansal M. A comprehensive guide to prompt engineering: unveiling the power of the COSTAR template [Internet]. Medium; 2024 Jan 10 [cited 2024 Apr 25]. <https://levelup.gitconnected.com/a-comprehensive-guide-to-prompt-engineering-unveiling-the-power-of-the-costar-template-944897251101>.

59. GovTech Data Science & AI Division. Prompt engineering playbook (Beta v3) [Internet]. Government of Singapore; 2023 Aug 30 [cited 2024 Apr 25]. <https://www.developer.tech.gov.sg/products/collections/data-science-and-artificial-intelligence/playbooks/prompt-engineering-playbook-beta-v3.pdf>.

60. Teo S. How I won Singapore's GPT-4 prompt engineering competition [Internet]. Medium; 2024 Dec 28 [cited 2024 Apr 25]. <https://towardsdatascience.com/how-i-won-singapores-gpt-4-prompt-engineering-competition-34c195a93d41>.

61. OpenAI. Prompt engineering [Internet]. OpenAI; [cited 2024 Apr 12]. <https://platform.openai.com>.

62. OpenAI, Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report [Internet]. arXiv; 2024 Mar 3 [cited 2024 Jul 16]. < http://arxiv.org/abs/2303.08774>.

DOI: dx.doi.org/10.5195/jmla.2025.1985

63. Vincent J. OpenAI co-founder on company's past approach to openly sharing research: "We were wrong" [Internet]. The Verge; 2023 Mar 15 [cited 2024 Jul 16]. <https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview>.

64. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus. 2023 May;15(5):e39238. DOI: https://doi.org/10.7759/cureus.39238

65. Giuse DA, Giuse NB, Bankowitz RA, Miller RA. Heuristic determination of quantitative data for knowledge acquisition in medicine. Comput Biomed Res Int J. 1991 Jun;24(3):261–72. DOI: https://doi.org/10.1016/0010-4809(91)90048-2

66. Kong A, Barnett GO, Mosteller F, Youtz C. How medical professionals evaluate expressions of probability. N Engl J Med. 1986 Sep 18;315(12):740–4. DOI: https://doi.org/10.1056/nejm198609183151206

67. Giuse DA, Giuse NB, Miller RA. Evaluation of long-term maintenance of a large medical knowledge base. J Am Med Inform Assoc. 1995 Sep 1;2(5):297–306. DOI: https://doi.org/10.1136/jamia.1995.96073832

68. Giuse NB, Giuse DA, Miller RA, Bankowitz RA, Janosky JE, Davidoff F, Hillner BE, Hripcsak G, Lincoln MJ, Middleton B. Evaluating consensus among physicians in medical knowledge base construction. Methods Inf Med. 1993 Apr;32(2):137–45.

69. Internet Archive. Wayback Machine [Internet]. Internet Archive; [cited 2024 Jul 16]. <https://web.archive.org/>.

70. Jerome RN, Giuse NB, Gish KW, Sathe NA, Dietrich MS. Information needs of clinical teams: analysis of questions received by the Clinical Informatics Consult Service. Bull Med Libr Assoc. 2001 Apr;89(2):177–84.

71. Random.org [Internet]. [cited 2024 Jul 16]. <https://www.random.org/>.

72. Zhang G, Jin Q, Jered McInerney D, Chen Y, Wang F, Cole CL, Yang Q, Wang Y, Malin BA, Peleg M, Wallace BC, Lu Z, Weng C, Peng Y. Leveraging generative AI for clinical evidence synthesis needs to ensure trustworthiness. J Biomed Inform. 2024 May;153:104640. DOI: https://doi.org/10.1016/j.jbi.2024.104640

73. Jerome RN, Giuse NB, Rosenbloom ST, Arbogast PG. Exploring clinician adoption of a novel evidence request feature in an electronic medical record system. J Med Libr Assoc JMLA. 2008 Jan;96(1):34–41. DOI: https://doi.org/10.3163/1536-5050.96.1.34

74. Piwowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, Farley A, West J, Haustein S. The state of OA: a large-scale analysis of the prevalence and impact of open access articles. PeerJ. 2018 Feb 13;6:e4375. DOI: https://doi.org/10.7717/peerj.4375

75. STM. Uptake of open access [Internet]. STM; [cited 2024 Apr 25]. <https://www.stm-assoc.org/oa-dashboard/uptake-of-open-access/>.

76. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. DOI: https://doi.org/10.3389/frai.2023.1169595

77. Topaz M, Peltonen LM, Michalowski M, Stiglic G, Ronquillo C, Pruinelli L, Song J, O'Connor S, Miyagawa S, Fukahori H. The ChatGPT effect: nursing education and generative artificial intelligence. J Nurs Educ. 2024 Feb 5:1-4. DOI: https://doi.org/10.3928/01484834-20240126-01

78. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, Aziz S, Damseh R, Alabed Alrazak S, Sheikh J. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. 2023 Jun 1;9:e48291. DOI: https://doi.org/10.2196/48291

79. Navigli R, Conia S, Ross B. Biases in large language models: origins, inventory, and discussion. J Data Inf Qual. 2023 Jun 22;15(2):10:1-10:21. DOI: https://doi.org/10.1145/3597307

80. Dorr DA, Adams L, Embí P. Harnessing the promise of artificial intelligence responsibly. JAMA. 2023 Apr 25;329(16):1347–8. DOI: https://doi.org/10.1001/jama.2023.2771

81. OpenAI. Introducing GPTs [Internet]. OpenAI; 2023 Nov 6 [cited 2024 Apr 25]. <https://openai.com/blog/introducing-gpts>.

82. Consensus: AI search engine for research [Internet]. Consensus; [cited 2024 Apr 25]. <https://consensus.app/>.

83. Scopus AI: Trusted content. Powered by responsible AI. [Internet]. Elsevier [cited 2024 Apr 25]. <https://www.elsevier.com/products/scopus/scopus-ai>.

84. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA. 2023 Sep 5;330(9):866–9. DOI: https://doi.org/10.1001/jama.2023.14217

85. Trinquart L, Johns DM, Galea S. Why do we think we know what we know? A metaknowledge analysis of the salt controversy. Int J Epidemiol. 2016 Feb;45(1):251–60. DOI: https://doi.org/10.1093/ije/dyv184

AUTHORS' AFFILIATIONS

Mallory N. Blasingame, MA, MSIS, mallory.n.blasingame@vumc.org, https://orcid.org/0000-0003-0356-9481, Information Scientist & Assistant Director for Evidence Provision, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Taneya Y. Koonce, MSLS, MPH, taneya.koonce@vumc.org, https://orcid.org/0000-0002-4014-467X, Deputy Director, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Annette M. Williams, MLS, annette.williams@vumc.org, https://orcid.org/0000-0002-2526-3857, Senior Information Scientist and Associate Director for Metadata Management, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Dario A. Giuse, Dr. Ing., MS, FACMI, dario.giuse@vumc.org, https://orcid.org/0000-0002-2677-6734, Associate Professor, Department of Biomedical Informatics, Vanderbilt University School of Medicine and Vanderbilt University Medical Center, Nashville, TN

DOI: dx.doi.org/10.5195/jmla.2025.1985

Jing Su, MD, MS, MSIS, jing.su@vumc.org, https://orcid.org/0000-0001-6699-6806, Senior Information Scientist, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Poppy A. Krump, poppy.krump@vumc.org, https://orcid.org/0000-0002-3081-6487, Information Scientist, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Nunzia Bettinsoli Giuse, MD, MLS, FACMI, FMLA, nunzia.giuse@vumc.org, https://orcid.org/0000-0002-7644-9803, Professor of Biomedical Informatics and Professor of Medicine; Vice President for Knowledge Management; and Director, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Received May 2024; accepted August 2024

ISSN 1558-9439 (Online)

CASE REPORT
DOI: dx.doi.org/10.5195/jmla.2025.1964

Using expert knowledge and peer review to create a reproducible process for the NAHRS Nursing Essential Resources List (NNERL)

Rebecca Raszewski, AHIP; Lorraine Porcello; Alissa V. Fial; Carolyn C. Dennison, AHIP; Rachel Keiko Stark, AHIP; Karen S. Alcorn; Sarah Wade

See end of article for authors' affiliations.

Background: Librarians have relied on resource lists for developing nursing collections, but these lists are usually in static or subscription-based formats. An example of this is the 26th edition of the Essential Nursing Resources last published in 2012. The Nursing and Allied Health Resources and Services (NAHRS) Caucus Nursing Essential Resources List (NNERL) Task Force has been working on a new list since Fall 2020. The goal of the Task Force is to create a nursing resource list that represents current materials and formats, uses a selection process that is transparent and reproducible, and will be available to a broad audience.

Case Presentation: Working from the Essential Nursing Resources 26th edition, the NNERL Task Force updated the purpose statement then began reviewing the resources on the list. Two working groups were formed: 1) an evaluation rubric working group developed a tool to evaluate the resources and 2) a tagging work group developed guidelines for creating metadata and "tags." Volunteers were recruited from the NAHRS Caucus to tag the resources. Lastly, the Task Force finalized the list of resources in the NNERL then cleaned and reconciled the data.

Conclusions: The final version of the NNERL will be published in Airtable, a cloud-based project management product, that will include metadata for every item on the list. The NNERL will be copyrighted to the NAHRS NNERL Task Force and made available through the Open Science Framework (OSF) under an Attribution-NonCommercial-NoDerivatives 4.0 International Creative Commons License.

Keywords: Library Collection Development; Libraries; Nursing; Case Reports; Nursing and Allied Health Resources and Services (NAHRS)

See end of article for supplemental content.

BACKGROUND

Resource lists of books, journals, or websites for the nursing profession have existed for decades. However, until now, these lists were either one-time publications, no longer updated, or required a paid subscription [1-6]. Librarians are often involved in creating these lists, using their professional knowledge to ensure the recommended materials are current and of high quality. Recommended resource lists help healthcare students and professionals find reliable and trusted resources for their studies, practice, or scholarship [7, 8]. While these lists have been helpful,, there is no evidence in the literature they are created in a reproducible and peer reviewed manner.

This case report describes the processes developed by a task force from the Nursing and Allied Health Resources and Service (NAHRS) Caucus of the Medical Library Association (MLA) to create the NAHRS Nursing Essential Resource List, hereafter called the NNERL. The processes included developing a rubric-based evaluation of nursing information resources, a vocabulary of metadata "tags" to describe each resource, and a structured review procedure. Through these improvements, the NNERL is a new resource that is accessible to all, regardless of professional membership. This case report also provides a model for similar groups within MLA to create or sustain collection development resources in other health sciences disciplines. By peer reviewing the materials included in the list and creating a reproducible methodology for the creation and maintenance of this list, the NNERL Task Force seeks to create a high-quality resource that follows current standards for open science, while being easier to update and maintain than traditional lists.

DOI: dx.doi.org/10.5195/jmla.2025.1964

The Essential Nursing Resources List

In 1966, the Interagency Council on Information Resources in Nursing (ICIRN) first published the Essential Nursing Resources List (ENRL) [9]. Expert information professionals with experience and responsibilities in nursing librarianship created the ENRL as a curated list of core resources for nursing libraries to utilize for collection development, current awareness, professional education, and career advancement. Schnall and Fowler published the 26th and final edition of the list in 2012 [10]. The ICIRN disbanded in 2017.

In October 2020, the NNERL Task Force leader recruited a team of volunteers from NAHRS to continue the ICIRN's work. Rather than follow the previous format and workflow, the Task Force aimed to reimagine the list as a living document and develop consistent methods for future updates. The Task Force leader charged two working groups to accomplish these goals. One group of four members focused on creating a rubric for evaluating resources, while the other group of six members created metadata "tags" with associated definitions to describe the NNERL resources. Completing their respective charges, each working group created supportive documentation and reproducible methodology for evaluating, indexing, and adding or removing resources from the NNERL.

CASE PRESENTATION

Updating Purpose Statement

The 2012 edition of the ENRL included 396 items available through print, electronic, and mobile formats as a static list. It was "presented as a resource for locating nursing information and for collection development… to support nursing practice, education, administration, and research activities. The list was compiled to point to pathways for exploration, rather than be an endpoint, and to expand to other formats beyond traditional references" [10]. Blogs, forums, and discussion lists were added to this version. Due to the issues with the permanence of blogs, forums, and discussion lists, as well as evolving professional standards, these materials were removed from this version of the NNERL.

To reflect the extensive changes made to the 2012 ENRL and to acknowledge the role of NAHRS in taking over the list, the 2024 NNERL is considered a new version. However, the purpose statement remained the same, except for adding language that the NNERL is for information professionals, and items "on this list represent high quality, evidence-based, and/or peer-reviewed resources appropriate for use in scholarly work, clinical practice, and research" [11]. The NNERL Task Force added an Attribution-NonCommercial-NoDerivatives 4.0 International Copyright designation (CC BY-NC-ND 4.0), enabling list users to copy or redistribute the NNERL and its associated materials [12].

Reviewing and Updating Resources

To begin, a Task Force member assigned each resource a numerical identifier and then moved all extant resources to a master spreadsheet. Next, each NNERL Task Force member received a subset of resources divided into blocks of roughly 25, with instructions provided for the initial review. Upon completion, each reviewer received a new set of resources until each resource was independently reviewed by at least two Task Force members. The instructions for the initial review included a color-coding system to indicate that a resource should be retained (green), removed (red), or required further discussion with the Task Force (yellow). During virtual monthly meetings, members discussed all resources that did not receive matching recommendations from the two reviewers, until achieving consensus.

In general, the Task Force's review of existing websites on the list resulted in retaining only the main website and removing the subsidiary sites. For example, the 2012 ENRL included the Centers for Disease Prevention and Control (CDC) homepage and links to subsidiary sites within the CDC.gov domain (e.g., CDC Wonder). Since website domain information deteriorates faster over time than information in other formats, the reduction of website subsidiaries aimed to minimize the impact of URL changes, which were assumed less likely for prominent organizations such as the CDC or WHO. For websites already on the list, only rarely and by consensus were subsidiary web resources retained.

Rubric Creation

To ensure the reproducibility of the peer review process, the rubric working group created a document with 9 questions that addressed current needs and expectations for inclusion in the NNERL when applied to each resource. This rubric would be used to evaluate the materials for inclusion to reduce the likelihood the Task Force would inject bias during the evaluation process of the included resources. The Task Force recommended using literacy standards and the NNERL's updated purpose statement as a starting point for the working group, both of which informed the rubric's creation. The above-listed standards included resources on evaluating information and internet resources [13-15]. The NNERL's updated purpose statement guided the resource evaluation criteria. Each rubric characteristic included scope notes (i.e. brief definitions and/or examples for use), creating a shared understanding of terms. The rubric designated the following characteristics to evaluate each resource:

- Authors or creators of listed resource
- Transparency of methods for resource creation
- Expertise of authors or creators
- Date of creation and/or last update

DOI: dx.doi.org/10.5195/jmla.2025.1964

- Frequency of updates
- Funding disclosure
- Conflict of interest disclosure
- Inclusion of historically underrepresented groups
- Relevance of the resource to health information professionals or a health sciences library collection

The rubric and the scope notes are available in Appendix A.

The NNERL Task Force used a small subset of NNERL resources to test the usability of the evaluation rubric. Four testers were recruited from within the Task Force for the first round of revisions, with three testers recruited from the larger NNERL Task Force for the second round of testing. To test the rubric, a small random selection of resources was assigned to testers, with each item being reviewed using the rubric twice, each time by a different tester. After the two reviewers' results were compared to ensure the rubric's usability, comments were solicited from all the testers. The rubric, once tested and completed, was then used by the NNERL Task Force to review each resource within the NNERL list. Each item in the NNERL list was reviewed by at least two different reviewers, and the results of those reviews were then used as part of the discussion for inclusion for each resource. Scores alone were not used to automatically accept or reject a resource. However, the scores collected for each item in the NNERL list will remain unpublished because they are not validated and only meaningful to the Task Force.

Tagging and Scope Note Creation

A second NNERL working group focused on creating "tags" for the items included on the NNERL list. The NNERL Task Force used the term "tag" to refer to the metadata assigned to individual resources. "Tag" was chosen as the descriptive term that the working group used due to lack of better options, there is no connection to the use of "tag" in this document with the kind of tagging created in social media, hence why "tag" is in quotations. The goal of creating "tags" as a form of metadata was to make items on the list searchable. While the ultimate goal of the NNERL Task Force is to make the list available online, which would require metadata creation, having "tags" available for each item in the list in its current form allows users to use a find option for materials based on their "tags." The working group developed a list of "tags" and a set of instructions to guide the process of assigning them. Then the "tags" were given to NNERL Task Force members to test by applying them to a random selection of resources from the NNERL list. Each item was reviewed by two volunteers, with feedback solicited from testers and modifications made based on feedback. The final list of "tags" included those from the 2012 ENRL, and newly created "tags" to reflect current trends, formats, and use. Next, the working group organized the "tags" into four categories: format, cost, areas of interest, and archival. During the review, some format tags from the ENRL were removed for being too specific (e.g., dictionaries); other format "tags" were added to assist librarians with their collection development and management responsibilities. For example, cost "tags" in the ENRL were previously limited to "Fee Required." The "tag," "Free," was added along with a new "tag," "Freemium," which aimed to identify resources that offer a "Free" version as well as a "Fee" version that includes more robust features and content for an additional cost. Areas of Interest "tags" (e.g., "Evidence-Based Nursing," "Informatics") from the previous edition were reviewed by the entire NNERL Task Force, with "tags" added and removed to reflect changes in the creation and dissemination of information. The Archival category of "tags" identifies resources that are no longer updated but have historical importance or significance for librarians (e.g., "multivolume sets," "bibliographies"). Table One lists the major descriptive "tag" categories and their definitions. The full list of "tags" and their scope notes are available in Appendix B.

Table 1 Descriptive "Tags"

Category	Definition
Format	A version of the item in which it is available (e.g., book, database, journals, serials, web-based resources). A Format "tag" is applied to every item.
Cost-Related to the Item	Use for item related costs: free, subscription-based, or freemium. Freemium is used for a free version with robust features and content for an additional fee. The Cost related to the item "tag" is applied to every item.
Areas of Interest/Topics	Descriptive terms are used to identify the subject content of each item. These "tags" are developed from MeSH (Medical Subject Headings) definitions and CINAHL (Cumulative Index of Nursing and Allied Health Literature) headings. Up to five "tags" are applied to every item.
Archival	Special "tag" only applied to items that are no longer updated.

A two-member team of the tagging working group created scope notes for each Area of Interest and Format "tag." Another two-member team of the tagging working group reviewed the scope notes, and the two teams

DOI: dx.doi.org/10.5195/jmla.2025.1964

discussed disagreements in wording until reaching a consensus. Each team met virtually, with the two teams meeting together three times to finalize the scope notes. Once the tagging working group completed the scope notes for the "tags," they developed instructions for assigning them. The guideline included in-depth information about each category and detailed scope notes for each Area of Interest "tag." The Task Force leader solicited volunteers from the NAHRS Caucus to test the guidelines for face validity [16] and inter-rater reliability. Each resource had two testers. The "tags" and the guidelines were adjusted based on the testers' feedback. For example, the testers suggested adding more "tags," particularly for areas of interest that were not part of the previous list (e.g., Multidisciplinary).

Final Review of Resources and Data

Spreadsheets were used in the development stage of the NNERL to ensure all members had access to the materials used by the group; spreadsheets were accessible to all members regardless of institutional limitations; however, the Task Force did not want the NNERL to remain in a spreadsheet format.

After the resources were scored and tagged, each resource's metadata was consolidated into one spreadsheet to clean and reconcile the data. NNERL Task Force members collaborated on updating the metadata and writing descriptions for each resource. They marked resources that were not deemed as essential to nursing for discussion by the entire Task Force during their monthly meetings. Several resources were removed during this process (e.g. Drugs@FDA, National Science Foundation), and a few more were added based on feedback obtained during reviewing the resources' "tags" (e.g. Nursing: Scope and Standards of Practice and Scopus). Next, the team selected which information would remain in the public version, and which would be archived such as raw rubric scores for resources. Finally, they standardized the data in each column to ensure consistency between entries. Once the spreadsheet was clean and usable, the "tags" were incorporated into the spreadsheet and reviewed.

Designing End Product

The Task Force's work on the NNERL incorporated best practices in collection development, assessment, and data management. Previous sections of this case report demonstrated the attention paid to choosing and evaluating information resources essential to supporting the work of nurses and nursing librarians. Of equal importance are the tenets of data management. Kipps & Jones point out that librarians routinely support researchers in designing research data management plans; however, "there are other data that are crucial to the library workplace. These include usage data, quantitative data . . . financial data, circulation data, and more" [17]. The NNERL project is a rich source of factual and descriptive metadata.

As such, NNERL Task Force members with experience in database creation and management explored options for transforming the cleaned list from a static spreadsheet to a searchable database. They focused on solutions that would address such requirements as an accessible and searchable interface, adequate data storage capacity, scalability to accommodate future growth of the list, robust tutorials, or support information, at little to no cost. Unsurprisingly, resource-intense tools such as a Structured Query Language (SQL) server to access data needed to meet additional requirements, so a cloud-based solution became the focus of the investigation. Cloud-based project management options that are available today combine data storage with functionality essential for teams whose members collaborate remotely. One such out-of-the-box solution is Airtable, which combines the features of spreadsheets with database functionality but does not require extensive prior knowledge to use efficiently [17]. More importantly, Airtable uses a freemium pricing model wherein the cost is not incurred until the database grows beyond 1,000 records. While it is notable that certain interface customizations are not available on the no-cost plan, Airtable is the current solution for the NNERL Task Force because of its data visualization and project management features.

Although an Airtable database, known as a "base" [18], can consist of many tables, the NNERL base currently has only one table. Future growth areas include adding tables for data related to NNERL. For example, when items are eventually removed from the public-facing table, they can be archived in a separate table accessible by designated individual(s) within the NAHRS Caucus.

Two team members initiated the transition from an Excel spreadsheet to Airtable base. They entered a subset of 60 records from the master spreadsheet into a database to learn how to use Airtable and test the usability of the spreadsheet data. The spreadsheet's columns are fields in the database, and each resource on the list is a record. The efforts invested in cleaning the data while still in spreadsheet format yields benefits because there is consistency between records for each field in the table. Plans to create visualizations of the NNERL records have been initiated beyond what is available in the grid-only view of a spreadsheet. For example, Airtable provides a "kanban view" that groups items into stacks of information cards based on a specified field [19]. Thus, grouping records by fields, such as format or cost, takes minutes. Customizing the visualization of NNERL records involves grouping, sorting, and color-coding records and fields but can also include hiding or filtering data [20]. For example, the rubric scores for each resource are not independently validated; as such, they are meaningful only to the Task Force. Filtering out these fields from the

DOI: dx.doi.org/10.5195/jmla.2025.1964

public view does not delete the data; but prioritizes the information needed for collection development.

A link to the NNERL through Airtable will be posted on the Open Science Framework (OSF) repository. A static version of the NNERL will also be provided on OSF for those working in settings with strict firewalls and network security.

DISCUSSION

The NNERL is an updated list of peer-reviewed and high-quality resources that can be used to develop solid and usable collections of nursing resources. By making the list freely available, information professionals, nurses, and those who work at nursing schools worldwide can rely on the NNERL to find appropriate resources. This case report represents the first step in the dissemination process, to be followed by communication to multiple health science librarian newsletters and listservs. Raising awareness of the NNERL with health sciences librarians in academic and hospital settings enables them to share it with their nursing faculty, staff nurses, and clinicians. Furthermore, targeted submissions to newsletters of nursing organizations such as the American Nursing Association and other nursing specialty organizations, will broaden the reach of the list.

Limitations

As with any task force, relying on unpaid volunteers can be challenging. Some librarians initially interested in working on the NNERL left the Task Force due to other time commitments. The final group of Task Force members were in different time zones across the United States (including Hawaii), so they were limited to when they could meet. Despite the challenges inherent in online collaboration across six US time zones, the Task Force assigned work to individuals or two-person teams who would provide progress reports during monthly meetings on Zoom. Eventually, the workflow evolved to include separate resource review meetings; although not required for the whole group, anyone working on assigned tasks could use these meetings to ask questions, resolve issues with specific resources, and determine the next steps. These meetings always generated rich discussions on collection development for nursing collections and which resources were considered essential.

The COVID-19 pandemic was a significant barrier for the work of the Task Force. Members tried to complete tasks alongside their primary job responsibilities, while adjusting to working remotely. Scoring print resources was particularly challenging because of multiple COVID-19 facility closures, which contributed to further delays in reviewing resources. Finding an online collaborative tool for working on documents together is an area for improvement. Although the Task Force selected Microsoft Teams, most members working remotely were using personal devices that did not allow access to Teams. Thus, team members received individual spreadsheets by email to update metadata or provide consolidated rubric scores. Merging multiple spreadsheets significantly complicated the data-cleaning process.

Future Plans/Moving Forward

After constructing the public interface, the next step is to create data entry forms using configuration tools available in Airtable. Once these tasks are completed, the NNERL is ready to succeed as the newest version of this valuable collection development tool. The commitment of the Task Force to reproducibility and transparency allows a documented framework for future groups to update the NNERL, which will maintain its quality and usability.

The NNERL Task Force developed a Leadership Position Description for a NAHRS Caucus member to serve as the overseer of the list. The expectation is the individual will serve in the role for two years. During that time, the individual will be a contact person for the Airtable platform, and quarterly elicit feedback on items currently in the list and additional items to be added. They will also seek volunteers to coordinate these efforts; utilizing the process developed by the Task Force. In addition, the Task Force will furnish NAHRS Caucus leadership with documentation for maintaining and growing the NNERL in years to come.

ACKNOWLEDGMENTS

The authors want to acknowledge fellow Task Force members Gary Atwood, Systematic Review Librarian & Library Associate Professor at University of Vermont; Basia Delawska-Elliott, Medical Librarian at Providence Health and Services; Diane Kearney, Lead Medical Librarian at Northern Virginia Community College; and Kristyn Seeman, Clinical Perioperative Research Librarian at AORN. The authors also acknowledge librarians Tess Grynoch, Research Data and Scholarly Communications Librarian at University of Massachusetts Chan Medical School; for her advice on metadata and Rachel Lerner, Reference & Instruction Librarian at Quinnipiac University; for her work on developing and consolidating the individualized spreadsheets used by the Task Force, and consultation on the transition to Airtable.

Appendix C acknowledges the librarians who were the original task force members or volunteered to tag the resources.

DATA AVAILABILITY STATEMENT

The rubric, "tags," a link to the NNERL in Airtable, and a static version of the NNERL will be posted on the OSF repository: https://osf.io/urjyz/?view_only=d3f4e0f57b864ee4838ae66b95eac858.

DOI: dx.doi.org/10.5195/jmla.2025.1964

AUTHOR CONTRIBUTIONS

Rebecca Raszewski: Conceptualization; Data Curation; Methodology; Project Administration; Supervision; Writing – Original Draft; Writing – Review & Editing; Lorraine Porcello: Conceptualization; Data Curation; Methodology; Resources; Visualization; Writing – Original Draft; Writing – Review & Editing; Alissa V. Fial: Conceptualization; Data curation; Methodology; Formal analysis; Supervision; Writing – Original Draft; Writing – Review & Editing; Carolyn C. Dennison, MA, MLIS, AHIP: Data curation; Methodology; Writing - review & editing; Rachel Keiko Stark: Conceptualization, Methodology, Writing- Original Draft, Writing - Review & Editing; Karen S. Alcorn: Conceptualization; data curation; methodology; formal analysis; resources; writing - original draft; writing - review & editing; Sarah Wade: Writing - Review & Editing; Data Curation; Formal Analysis

REFERENCES

1. Jacobs SK, Bertolucci Y; Interagency Council on Information Resources in Nursing. Essential nursing resources. Nurs Educ Perspect. 2007 Sep-Oct;28(5):276-85.
2. Sherwill-Navarro P, Kennedy JC, Allen MP. Developing an evidence-based list of journals for nursing. J Med Libr Assoc. 2014 Apr;102(2):105-9. https://doi.org/10.3163/1536-5050.102.2.009.
3. Lomax E, Setterlund SK. A virtual reference shelf for nursing students and faculty. J Libr Adm. 2008 Oct 12;44(1-2):429-51. https://doi.org/10.1300/J111v44n01_06.
4. Owens JW. Nursing tools for the selection of library resources. Acquis Libr. 2009 Jan 20;16(31–32):89–99. https://doi.org/10.1300/J101v16n31_08.
5. Hill DR, Stickell HN. Brandon/Hill selected list of print nursing books and journals. Nurs Outlook. 2002 May-Jun;50(3):100-13. https://doi.org/10.1067/mno.2002.126030.
6. Doody's Core Titles. Wolters Kluwer. https://www.doody.com/dct/.
7. Clark TR. Resource list on medication use in older adults. Generations. 2011;35(4):83-4. https://www.jstor.org/stable/26555904.
8. Moffitt, C. Creating a comprehensive blindness and low vision resource list. J Vis Impair Blind. 2022 Jul;116(4):563-6. https://doi.org/10.1177/0145482X221121671.
9. Committee of the Interagency Council on Library Tools for Nursing. Reference tools for nursing. Nurs Outlook. 1966 May;14(5):67-72.
10. Schnall JG, Fowler S. (2012). Essential nursing resources: one source for evidence-based nursing practice. Am J Nurs. 2012 Jan;112(1):23. https://doi.org/10.1097/01.NAJ.0000410354.34028.c3.
11. Raszewski R, Stark RK, Learner RC, Alcorn K, Dennison CC, Fial A, Kearney D, Wade S, Delawska-Elliott BB. Forging ahead with the NAHRS nursing essential resources list: utilizing expert knowledge and peer review processes to create a reproducible resource and workflow. Presented at: MLA '23, the 123rd Medical Library Association Annual Meeting; Detroit, MI; May 19, 2023.12
12. Creative Commons. https://creativecommons.org/licenses/by-nc-nd/4.0/.
13. Johns Hopkins Sheridan Libraries. Evaluating information [Internet]. John Hopkins University; 2023 [cited 5 Mar 2024]. https://guides.library.jhu.edu/evaluate/sources.
14. Medical Library Association. Find good health information [Internet]. Medical Library Association [cited 5 Mar 2024]. https://www.mlanet.org/resources/userguide.html.
15. Penn State University Libraries. Evaluating information rubric [Internet]. Penn State University [cited 5 Mar 2024]. https://libraries.psu.edu/research/how/evaluating-information-rubric.
16. Tanner K. Survey designs. In: Williamson K, Johanson G. Research Methods: Information, Systems, and Contexts. 2nd ed. United Kingdom: Chandos Publishing; 2018:159-92. https://doi.org/10.1016/B978-0-08-102220-7.00006-6.
17. Kipps K, Jones AK. Collection management in the cloud: a guide for using cloud computing technologies in libraries. Blue Ridge Summitt, PA: Rowman & Littlefield Publishers; 2022:37-8.
18. Airtable. https://www.Airtable.com/guides/build/create-a-base.
19. Airtable. https://support.Airtable.com/docs/getting-started-with-Airtable-kanban-views.
20. Airtable. https://www.Airtable.com/guides/build/create-custom-views-of-data.

SUPPLEMENTAL FILES

- **Appendix A**: Rubric
- **Appendix B:** "Tags" with scope notes
- **Appendix C:** Additional Acknowledgements

AUTHORS' AFFILIATIONS

Rebecca Raszewski, MS, AHIP, raszewr1@uic.edu, https://orcid.org/0000-0003-1210-4272, University of Illinois Chicago, Library of the Health Sciences Chicago, Chicago, IL

Lorraine Porcello, MSLIS, MSIM, Lorraine_Porcello@URMC.Rochester.edu, https://orcid.org/0000-0002-6077-6631, University of Rochester, Rochester, NY

Alissa V. Fial, MA, MLIS, alissa.fial@marquette.edu, https://orcid.org/0000-0001-6927-434X, Marquette University, Milwaukee, WI

Carolyn C. Dennison, MA, MLIS, AHIP, cdenniso@hawaii.edu, University of Hawaii at Manoa, Honolulu, HI

DOI: dx.doi.org/10.5195/jmla.2025.1964

Rachel Keiko Stark, MS, AHIP, stark@csus.edu, https://orcid.org/0000-0001-7946-2213, Sacramento State University, Sacramento, CA Karen S. Alcorn, MLS, karen.alcorn@mcphs.edu, https://orcid.org/0000-0002-3184-0665, Massachusetts College of Pharmacy and Health Sciences, Worcester, MA

Sarah Wade, MLS, swade@campbell.edu, https://orcid.org/0000-0002-7640-6715, Campbell University School of Osteopathic Medicine, Lillington, NC

Received May 2024; accepted September 2024

ISSN 1558-9439 (Online)

DOI: dx.doi.org/10.5195/jmla.2025.2102

JMLA virtual projects continue to show impact of technologies in health sciences libraries

Emily Hurst, AHIP

See end of article for authors' affiliations.

Beginning in 2012, the Virtual Projects section of the *Journal of the Medical Library Association* has provided an opportunity for library leaders and technology experts to share with others how new technologies are being adopted by health sciences libraries. From educational purposes to online tools that enhance library services or access to resources, the Virtual Projects section brings technology use examples to the forefront. The new publication issue for future Virtual Projects sections will be January and the call for submissions and Virtual Projects deadline will now take place in June and July.

Virtual Projects are published on an annual basis in the *Journal of the Medical Library Association* (*JMLA*) following an annual call for virtual projects in *MLAConnect* and announcements to encourage submissions from all types of libraries. An advisory committee of recognized technology experts selects project entries based on their currency, innovation, and contribution to health sciences librarianship.

The Virtual Projects Section continues to evolve, enlighten, and offer a forum for health sciences libraries to share their ongoing work with technology-focused projects. This year the *JMLA* Virtual Projects Committee was delighted by the number of project abstracts received. Since 2020, those submitting content proposals for the Medical Library Association (MLA) conference may opt in to have their abstract reviewed by the *JMLA* Virtual Projects Committee for consideration in the Virtual Projects Section. This opportunity has opened the door, allowing even more projects to be considered.

This year the Virtual Projects Committee is pleased to share eight projects that demonstrate the wide depth and breadth of technology deployment in health sciences libraries. A number of projects this year demonstrate early adoption of artificial intelligence (AI). From the use of AI to facilitate event planning to AI assistance in collection development as well as AI use in library research guide creation, AI is transforming nearly every aspect of health sciences librarianship. Also of note, this year several projects demonstrate the robust use of technology to enhance teaching and training including both in-person sessions and online modules. Unique ways to support the evolution and training of bioinformatics are also showcased in one project. Another project provides perspectives on the use of free and paid tools to quickly create and maintain clinical bibliographies and technology solutions to enhance repositories through the creation of digital object identifiers (DOIs) are highlighted in another.

Material for this year's column was selected in 2024 with the assistance of the *JMLA* Virtual Projects Advisory Committee: Christine Andresen; Emily J. Hurst section editor; Michelle Kraft, AHIP, section coeditor; J. Dale Prince, AHIP; Tariq Rahaman; and Brian Zelip. Selected projects were edited by Emily J. Hurst.

AUTHORS' AFFILIATIONS

Emily J. Hurst, MSLS, AHIP, ehurst@hshsl.umaryland.edu, https://orcid.org/0000-0003-4191-1938, Virtual Projects Advisory Committee, Journal of the Medical Library Association, and University of Maryland Baltimore, Health Sciences and Human Services Library,

Received October 2024; accepted November 2024

ISSN 1558-9439 (Online)

VIRTUAL PROJECT
DOI: dx.doi.org/10.5195/jmla.2025.2076

Individual DOI minting for Open Repository: a script for creating a DOI on demand for a DSpace repository

Tess Grynoch; Lisa A. Palmer, AHIP

See end of article for authors' affiliations.

Digital Object Identifiers (DOIs) are a key persistent identifier in the publishing landscape to ensure the discoverability and citation of research products. Minting DOIs can be a time-consuming task for repository librarians. This process can be automated since the metadata for DOIs is already in the repository record and DataCite, a DOI minting organization, and Open Repository, a DSpace repository platform, both have application programming interfaces (APIs). Existing software enables bulk DOI minting. However, the institutional repository at UMass Chan Medical School contains a mixture of original materials that need DOIs (dissertations, reports, data, etc.) and previously published materials that already have DOIs such as journal articles.

An institutional repository librarian and her librarian colleague with Python experience embarked on a paired programming project to create a script to mint DOIs on demand in DataCite for individual items in the institution's Open Repository instance. The pair met for one hour each week to develop and test the script using combined skills in institutional repositories, metadata, DOI minting, coding in Python, APIs, and data cleaning. The project was a great learning opportunity for both librarians to improve their Python coding skills. The new script makes the DOI minting process more efficient, enhances metadata in DataCite, and improves accuracy. Future script enhancements such as automatically updating repository metadata with the new DOI are planned after the repository upgrade to DSpace 7.

Keywords: Institutional Repositories; DSpace; Open Repositories; DataCite; Python

Virtual Projects are published on an annual basis in the *Journal of the Medical Library Association* (*JMLA*) following an annual call for virtual projects in *MLAConnect* and announcements to encourage submissions from all types of libraries. An advisory committee of recognized technology experts selects project entries based on their currency, innovation, and contribution to health sciences librarianship.

BACKGROUND

The eScholarship@UMassChan institutional repository [1] is a digital archive and dissemination platform for the scholarship of students, faculty, and staff at UMass Chan Medical School in Worcester, Massachusetts. eScholarship@UMassChan utilizes Open Repository version 5.7, a hosted software platform from Atmire built on DSpace software [2]. The repository contains the full text of previously published items, such as journal articles, along with original materials, including theses, dissertations, posters, reports, and datasets.

For all original materials submitted, repository staff creates a Digital Object Identifier (DOI) via DataCite [3], a DOI minting organization. This is a crucial service, as DOIs are a key persistent identifier in the publishing landscape to ensure discoverability and citation of research products. DSpace repositories can mint DOIs automatically for all items, but this feature is not suitable for repositories that include published materials which already have DOIs. Thus, repository staff minted DOIs by entering metadata for each original resource into an online form, a time-consuming task open to error.

Knowing this process could be automated, the authors, a repository librarian and a data librarian colleague with coding experience, embarked on a paired programming project to create a Python script to mint DOIs on demand in DataCite for individual items in eScholarship@UMassChan.

THE PAIRED PROGRAMMING PROJECT

The repository librarian and data librarian met for one hour each week starting in July 2023 to develop the script. This approach took advantage of the librarians' combined skills in institutional repositories, metadata, DOI minting, coding in Python, APIs, and data cleaning. An important step was to create a crosswalk to map metadata values for document types from DSpace to DataCite. The project team also knew that the repository would be upgraded to DSpace 7 in 2024 and have a new API, leading them to only use the Open Repository API to download but not edit repository metadata. Another decision point was how

DOI: dx.doi.org/10.5195/jmla.2025.2076

to handle items with multiple authors with ORCID IDs, as repository metadata does not link authors with their ORCID. The project team decided that the script would utilize the ORCID field if the item had one author but not for multiple authors. So, items with multiple ORCID IDs still need to have IDs added to DataCite manually.

The script was successfully used to upload repository metadata to the DataCite test server in November 2023 and the production version of the script was created and tested in December 2023. A de-identified version of the script was published on GitHub in January 2024 and can be modified for use in other DSpace 5.7 repositories [4].

IMPACT AND FUTURE DIRECTIONS

The new script allows repository staff to mint DOIs more efficiently (3-13 minutes faster) with improved metadata and fewer human errors. Library users and the institution benefit because the institutional repository librarian has more time to enhance and add content to the repository. The project team also gained new skills that can be applied to additional opportunities to improve library processes and services to users. After the anticipated DSpace 7 upgrade, the project team plans to update the script using the new API. The project team is also monitoring DSpace/ORCID integration efforts that could improve the process [5].

AUTHOR CONTRIBUTIONS

Tess Grynoch: Conceptualization; Formal Analysis; Investigation; Methodology; Project Administration; Software; Resources; Validation; Writing – original draft; Writing – review & editing. Lisa Palmer: Conceptualization; Formal Analysis; Investigation; Methodology; Project Administration; Software; Resources; Validation; Writing – original draft; Writing – review & editing.

REFERENCES

1. eScholarship@UMassChan [Internet]. 2024 [cited 19 Sept 2024]. https://repository.escholarship.umassmed.edu/
2. Atmire Open Repository [Internet]. 2024 [cited 19 Sept 2024]. https://www.atmire.com/open-repository
3. DataCite [Internet]. 2024 [cited 19 Sept 2024]. https://datacite.org/
4. Mint DOI for Open Repository [Internet]. 2024 [cited 19 Sept 2024]. https://github.com/grynoch/mintDOI4OpenRepository,
5. Lyrasis. Lyrasis receives ORCID global participation fund grant to support ORCID integration improvements in DSpace [Internet]. 2024 March [cited 19 Sept 2024]. https://lyrasisnow.org/lyrasis-receives-orcid-global-participation-fund-grant-to-support-orcid-integration-improvements-in-dspace/ .

AUTHORS' AFFILIATIONS

Tess Grynoch, MLIS, Tess.Grynoch@umassmed.edu, https://orcid.org/0000-0003-0550-3961, Research Data & Scholarly Communications Librarian, Lamar Soutter Library, University of Massachusetts Chan Medical School, Worcester, MA

Lisa A. Palmer, MSLS, AHIP, Lisa.Palmer@umassmed.edu, https://orcid.org/0000-0002-4116-1279, Institutional Repository Librarian, Lamar Soutter Library, University of Massachusetts Chan Medical School, Worcester, MA

Received September 2024; accepted October 2024

ISSN 1558-9439 (Online)

VIRTUAL PROJECTS
DOI: dx.doi.org/10.5195/jmla.2025.2087

Leveraging AI tools for streamlined library event planning: a case study from Lane Medical Library

Boglarka Huddleston; Colleen Cuddy

See end of article for authors' affiliations.

Health sciences and hospital libraries often face challenges in planning and organizing events due to limited resources and staff. At Stanford School of Medicine's Lane Library, librarians turned to artificial intelligence (AI) tools to address this issue and successfully manage various events, from small workshops to larger, more complex conferences. This article presents a case study on how to effectively integrate generative AI tools into the event planning process, improving efficiency and freeing staff to focus on higher-level tasks.

Keywords: Artificial Intelligence; event planning; medical library

Virtual Projects are published on an annual basis in the *Journal of the Medical Library Association* (*JMLA*) following an annual call for virtual projects in *MLAConnect* and announcements to encourage submissions from all types of libraries. An advisory committee of recognized technology experts selects project entries based on their currency, innovation, and contribution to health sciences librarianship.

BACKGROUND AND CONTEXT

In the past year, Lane Medical Library organized a series of events using AI tools to assist with planning. Individual or small teams of librarians typically are responsible for event content and they work with the library marketing and communications team to promote events. Librarians are responsible for planning and executing events from start to finish. Once speaker(s), date, location, and event program are confirmed, the marketing team creates promotional materials that are distributed via different channels (newsletters, listservs, social media, etc.). The marketing and communications team is comprised of three staff members, led by and including the library's Web Services Librarian. Staff (librarians and/or library staff) rotate through the team on an annual basis.

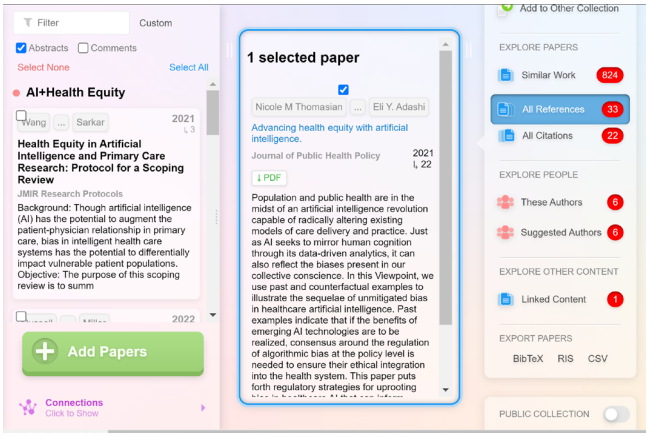
Recently, two notable events took place where event organizers heavily relied on the assistance of AI tools: 1) a live, online conversation with the co-founder of MedRx with 35 participants in attendance; and 2) a half-day online conference with over 250 participants titled "Women in Data Science: Artificial Intelligence and Health Equity". While these events differed in complexity, both benefited from AI support, which streamlined various aspects of the planning and execution process.

EVENT PLANNING WITH AI

Regardless of event size and complexity, the event planning process can be divided into three stages: pre-event, during the event, and post-event. Each stage offers unique opportunities to leverage AI tools to improve efficiency and reduce manual workload.

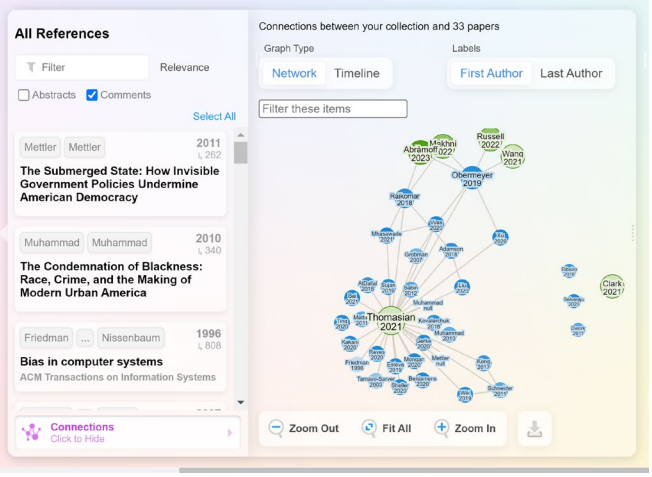
Pre-Event: During the pre-event phase, event organizers used AI tools to help manage content, structure planning, and marketing. For instance, Lane ResearchRabbit, an AI tool that allows users to search for papers and authors, monitor new literature, and visualize research landscapes, to identify potential speakers by generating interactive visualizations of relevant networks of papers and authors significantly reduced the time required to find and vet speakers for the events (see Figure 1 and Figure 2).

Figure 1 Selecting sample paper in ResearchRabbit



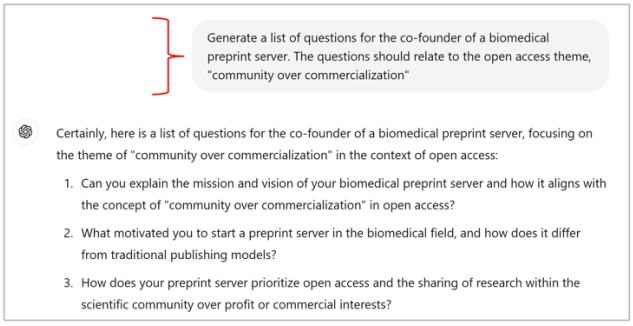
DOI: dx.doi.org/10.5195/jmla.2025.2087

Figure 2 Visualizing connections to relevant papers in ResearchRabbit



Additionally, event organizers employed ChatGPT 4.0 to create show-flow agendas, event pacing, speaker introductions, and panel questions, saving several manual work hours.

Figure 3 ChatGPT 4.0 Prompt for panel questions



During the Event: During live events, event organizers relied on ChatGPT 4.0 to generate on-the-fly panel discussion questions, allowing staff to focus on facilitating the event and engaging with participants.

Post-Event: After each event, ChatGPT 4.0 assisted in crafting follow-up communications, including thank-you emails to speakers and participants. The event organizers also used AI to generate post-event surveys, helping gather feedback to refine future events.

CONCLUSION

These experiences have demonstrated the benefits of generative AI tools for event planning, particularly in automating repetitive tasks and reducing time spent on event logistics, allowing organizers to focus on content creation and community engagement. However, event organizers also encountered a few challenges. AI-generated content sometimes included inaccuracies or overly complex language that required human intervention. In these cases, they had to manually revise speaker bios that contained incorrect information and streamline AI-generated email templates that contained superfluous content. This required fact-checking, for example, reviewing content against online profiles or LinkedIn to verify speaker information and some judicious editing for communications. Staff learned to create well-structured, clear AI prompts, through trial and error and sharing prompt libraries, which led to more accurate, relevant outputs, minimizing errors and misinterpretations. In the end, this improved efficiency, saved time, and enabled faster completion of tasks, thus enhancing overall productivity in the workflow.

AI tools have greatly enhanced event planning capabilities at Lane Medical Library, particularly in streamlining complex tasks and automating routine processes. While these tools require careful oversight and occasional adjustments, they have enabled library staff to deliver high-quality events more quickly and efficiently. For other libraries facing similar challenges, the Lane Library recommends exploring AI tools to augment traditional planning processes and maximize available resources.

AUTHORS' AFFILIATIONS

Boglarka Huddleston, MLIS, MA, b.huddleston@stanford.edu, Manager, Research & Instruction, Lane Medical Library, Stanford University School of Medicine, Stanford, MA

Colleen Cuddy, PhD, MLS, MA, ccuddy@stanford.edu, Director, Research and Academic Collaboration, Lane Medical Library, Stanford University School of Medicine, Stanford, MA

Received October 2024; accepted October 2024

ISSN 1558-9439 (Online)

VIRTUAL PROJECT
DOI: dx.doi.org/10.5195/jmla.2025.2085

Designing for impact: a case study of UTHSC's research impact challenge

Jess Newman McDonald; Annabelle L. Holt

See end of article for authors' affiliations.

Prompted by increasing requests for assistance with research evaluation from faculty researchers and university leadership, faculty librarians at the University of Tennessee Health Science Center (UTHSC) launched an innovative Research Impact Challenge in 2023. This Challenge was inspired by the University of Michigan's model and tailored to the needs of health sciences researchers. This asynchronous event aimed to empower early-career researchers and faculty seeking promotion and tenure by enhancing their online scholarly presence and understanding of how scholarship is tracked and evaluated.

A team of diverse experts crafted an engaging learning experience through the strategic use of technology and design. Scribe slideshows and videos offered dynamic instruction, while written content and worksheets facilitated engagement and reflection. The Research Impact Challenge LibGuide, expertly designed with HTML and CSS, served as the central platform, ensuring intuitive navigation and easy access to resources (https://libguides.uthsc.edu/impactchallenge). User interface design prioritized simplicity and accessibility, accommodating diverse learning preferences and technical skills.

This innovative project addressed common challenges faced by researchers and demonstrated the impactful use of technology in creating an adaptable and inclusive educational experience. The Research Impact Challenge exemplifies how academic libraries can harness technology to foster scholarly growth and support research impact in the health sciences.

Keywords: Education Technology; Research Metrics; Library Instructions; LibGuides; Health Science Libraries; Research Data Management; Outreach

Virtual Projects are published on an annual basis in the *Journal of the Medical Library Association* (*JMLA*) following an annual call for virtual projects in *MLAConnect* and announcements to encourage submissions from all types of libraries. An advisory committee of recognized technology experts selects project entries based on their currency, innovation, and contribution to health sciences librarianship.

The Research Impact Challenge was an asynchronous 5-day instructional series that ran August 7-11, 2023. Daily "Challenge" activities were designed to help participants better understand and manage their online scholarly presence and the impact and reach of their research. The intended audience for these activities was health science researchers and their support personnel, including graduate students.

- Day 1: Claim or Enhance Your Online Presence
- Day 2: Understand and Locate Your Research Impact Metrics
- Day 3: Reach a Wider Audience with Open Access Publishing
- Day 4: Learn How and Why to Share Your Research Data
- Day 5: Stay Informed and Build Your Network

The 2023 Challenge team comprised eight faculty librarians, one communications and marketing specialist, and one Institutional Research staff member. Three library faculty worked on the initial SpringShare LibGuide design, including one HTML/CSS specialist. Subsequent edits to the LibGuide content were made by the project lead, including updates for year two of the challenge.

The technology utilized throughout the challenge included a custom LibGuide, videos, written instructions, and Scribes to users to engage with the content on each day of the challenge. The custom LibGuide, designed with HTML and CSS, served as the central platform to facilitate intuitive navigation and seamless access to resources. The interface design prioritized simplicity and accessibility, accommodating a wide range of learning preferences and technical skills. For each daily topic, the project team used collapsible accordion navigation tools or menus to display the tasks involved to complete each challenge. Each task

DOI: dx.doi.org/10.5195/jmla.2025.2085

included a combination of written instructions, videos, links, and Scribes. The team employed the free version of Scribe to develop step-by-step tutorials, enabling users to complete tasks asynchronously with ease (https://scribehow.com/).

In 2023 the challenge had 62 total registrations, including 40 faculty, 16 staff, and five students. The LibGuide received 352 views during the challenge week, August 7-11, 2023. At the end of the week a Qualtrics feedback survey was distributed to participants. Participation was encouraged by offering a "Certificate of Completion" upon finishing the survey. All respondents (19 total) were affiliated with UTHSC and the majority were faculty (73%). Of those that did not complete the challenge (21%), most indicated that they intend to finish it as they have time. Seventy-nine percent indicated that they were highly likely (7 or above out of 10) to recommend the challenge to a colleague. Impact Metrics, Scholarly Profiles, and Data Sharing were the most highly rated days, although all days received above-average ratings. Participants strongly agreed that the Challenge was "well organized" and of "sufficient quality." Forty-two percent of respondents were surprised to learn that the Library had expertise on these topics, an additional 37% were somewhat surprised, indicating that the event was successful in promoting librarian expertise. Several participants, via this survey and personal email correspondence, expressed hope that the Challenge would remain available online for future use and dissemination.

For the second iteration of the Research Impact Challenge all content was created and updated in the LibGuide by the project lead. The challenge was again advertised and marketed by the communications specialist via various social media and internal communication channels. The 2024 challenge saw a slight drop in registrations, 57 total. The LibGuide received 185 views over the second challenge week, down from 352 the first year. This may be due to a variety of factors - a shorter lead up time for marketing in year two, the majority of the challenge content remained unchanged and thus there was no incentive for repeat participation, and prospective participants may have understood that registration was not required to access the content and that it would remain available after the challenge. Participants again received an anonymous link to a Qualtrics feedback survey with a new 2024 Certificate of Completion. The UT Health Science Center Library envisions that the challenge will be re-released annually with slight modifications and updates. As of Fall 2024 Library liaisons are in conversation with a College of Health Professions faculty member to include the challenge modules in their curriculum for the upcoming 2025 spring semester.

AUTHORS' AFFILIATIONS

Jess Newman McDonald, mcdo1364@umn.edu, https://orcid.org/0000-0001-5689-1656, University of Minnesota, MN

Annabelle L. Holt, annabelle.holt@uky.edu, https://orcid.org/0000-0003-4965-4235, University of Kentucky, KY

Received October 2024; accepted October 2024

ISSN 1558-9439 (Online)

VIRTUAL PROJECT
DOI: dx.doi.org/10.5195/jmla.2025.2079

Making the most of Artificial Intelligence and Large Language Models to support collection development in health sciences libraries

Ivan Portillo; David Carson

See end of article for authors' affiliations.

This project investigated the potential of generative AI models in aiding health sciences librarians with collection development. Researchers at Chapman University's Harry and Diane Rinker Health Science campus evaluated four generative AI models—ChatGPT 4.0, Google Gemini, Perplexity, and Microsoft Copilot—over six months starting in March 2024. Two prompts were used: one to generate recent eBook titles in specific health sciences fields and another to identify subject gaps in the existing collection. The first prompt revealed inconsistencies across models, with Copilot and Perplexity providing sources but also inaccuracies. The second prompt yielded more useful results, with all models offering helpful analysis and accurate Library of Congress call numbers. The findings suggest that Large Language Models (LLMs) are not yet reliable as primary tools for collection development due to inaccuracies and hallucinations. However, they can serve as supplementary tools for analyzing subject coverage and identifying gaps in health sciences collections.

Keywords: Generative artificial intelligence; large language models; ChatGPT; Microsoft Copilot; Perplexity; Google Gemini; collection development; collection assessment; health sciences libraries

Virtual Projects are published on an annual basis in the *Journal of the Medical Library Association* (*JMLA*) following an annual call for virtual projects in *MLAConnect* and announcements to encourage submissions from all types of libraries. An advisory committee of recognized technology experts selects project entries based on their currency, innovation, and contribution to health sciences librarianship.

Artificial intelligence (AI) and large language models (LLMs) have garnered significant interest since the public launch of ChatGPT in 2022 [1]. LLMs and generative AI models have made substantial strides in their capabilities, now offering references and detailed analyses of uploaded files in their responses. These advancements present a promising opportunity for librarians to potentially reduce workload and increase efficiency [2,3]. This project was designed to explore the potential of generative AI models in assisting health sciences librarians with collection development, particularly in identifying gaps and recommending book titles.

Chapman University is a private university with two campuses in Orange County, California, with approximately 10,000 students and 2,000 staff and faculty. The researchers are health sciences librarians based at the Harry and Diane Rinker Health Science campus in Irvine, CA, which serves primarily graduate and doctoral students in physical therapy, physician assistant, communication sciences, and pharmacy programs. Beginning in March 2024, the researchers evaluated four generative AI models over a period of six months using two prompts designed by the researchers to aid librarians in collection development. The four generative AI models assessed included ChatGPT 4.0, Google Gemini, Perplexity, and Microsoft Copilot.

The first prompt used in each generative AI model sought to generate a list of recent eBook titles published in the last two years focused on physical therapy, physician assistant, communication sciences and disorders, and pharmacy. The second prompt sought to identify subject gaps in an existing collection and create a list of recommended call number ranges. To accomplish this task, a list of the library's collection was uploaded into each generative AI model. The list was created using the Create List function within Sierra, an Integrated Library System from Innovative. The list was exported as a CSV file with fields for title, Library of Congress call number, location, and item status. If the AI model did not accept CSV files, such as the non-premium versions of Perplexity and Google Gemini, the researchers copied and pasted the list of titles and Library of Congress call numbers from the collection into the prompt field.

The results were assessed based on quality, accuracy, the presence of fabricated titles (often referred to as "hallucinations"), if references were provided, correct

DOI: dx.doi.org/10.5195/jmla.2025.2079

citation details, and accurate Library of Congress (LC) call numbers. Each AI model produced inconsistent results for the first prompt. Five titles per subject were generated by each AI model, with Copilot and Perplexity being the only two that provided sources for the titles generated. Perplexity generated inaccurate details, including publication years, DOIs, and publishers. Copilot was the most accurate, while Gemini and ChatGPT provided inaccuracies and hallucinations. It should be noted that the researchers found that all four AI models generated hallucinations and inaccurate information on previous dates with the same prompt provided. While Perplexity and Copilot performed the best, the researchers would not recommend any generative AI models for title recommendations due to inaccuracies and inconsistencies.

The researchers found the second prompt more helpful from each of the four generative AI models. Each AI model provided helpful analysis and accurate LC call numbers. Each AI model provided minor differences in the subject gaps they identified, but all provided the reasoning behind the importance of each subject area recommended. For example, when asked to identify subject gaps for physical therapy, ChatGPT and Copilot agreed on eight broad subject gaps, such as kinesiology and geriatrics. Perplexity and Gemini offered narrower, more specific suggestions, such as telehealth and electrophysical agents. The researchers have found this to be useful in their current collection development cycle.

Overall, the results reinforce the notion that LLMs are not yet suitable as primary information retrieval systems in collection development. It should be noted that the researchers found that all four LLMs generated hallucinations for prompt #1 and inaccurate information on previous dates. Responses also varied for prompt #2 depending on the day or time queried. The researchers still found that generative AI models can serve as a supplementary tool for analyzing the subject coverage of their collection, identifying subject gaps, and highlighting areas for health science programs that may not be as well represented in a library collection.

REFERENCES

1. Tian S, Jin Q, Yeganova L, Lai P-T, Zhu Q, Chen X, Yang Y, Chen Q, Kim W, Comeau DC, Islamaj R, Kapoor A, Gao X, Lu Z. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Briefings in Bioinformatics. 2024;25(1):bbad493. DOI: https://doi.org/10.1093/bib/bbad493.
2. Brzustowicz R. From ChatGPT to CatGPT: The Implications of Artificial Intelligence on Library Cataloging. Information Technology and Libraries. 2023;42(3). DOI: https://doi.org/10.5860/ital.v42i3.16295
3. Yamson GC. Immediacy as a better service: Analysis of limitations of the use of ChatGPT in library services. Information Development. 2023;0(0):02666669231206762. DOI: https://doi.org/10.1177/02666669231206762

AUTHORS' AFFILIATIONS

Ivan Portillo, iportillo@chapman.edu, https://orcid.org/0000-0002-8031-1006, Health Sciences Librarian, Director of Rinker Campus Library Services, Leatherby Libraries, Chapman University, Irvine, CA

David Carson, carsondav@ohsu.edu, https://orcid.org/0009-0002-6533-3159, Health Sciences Education & Research Librarian, Oregon Health & Science University, Portland, OR

Received October 2024; accepted October 2024

ISSN 1558-9439 (Online)

VIRTUAL PROJECT
DOI: dx.doi.org/10.5195/jmla.2025.2086

Leveraging an open access platform to provide organizational value in clinical environments

Aida Marissa Smith, AHIP; Alexia Estabrook; Mary A. Hyde, AHIP; Michele Matucheski, AHIP; Eleanor Shanklin Truex

See end of article for authors' affiliations.

The Ascension Nurse Author Index is an example of how resource-limited clinical libraries can provide value to their organization by creating a database of peer-reviewed journal article publications authored by their nursing associates. In 2024, Ascension launched a database index to highlight its nurse authors, bring attention to subject matter expertise, foster collaboration among authors, and recognize impact within the profession. The index uses an open access platform, software intended for reference management with a public-facing cloud option, to minimize expenses. This unconventional use of the platform allowed us to capitalize on the software's bibliographic database management capabilities while allowing us to input institutional-specific metadata. By creative use of the open-access platform, librarians can successfully partner to create value for their organization by highlighting the work of its nurses.

Keywords: Clinical Librarians, Hospital Librarians, Nurses, Bibliographic Management Software, Organizational Value, Authorship, Collaboration

Virtual Projects are published on an annual basis in the *Journal of the Medical Library Association* (*JMLA*) following an annual call for virtual projects in *MLAConnect* and announcements to encourage submissions from all types of libraries. An advisory committee of recognized technology experts selects project entries based on their currency, innovation, and contribution to health sciences librarianship.

The Ascension Nurse Author Index (Ascension, 2024) is an example of how resource-limited clinical libraries can provide value to their organization by creating a database of peer-reviewed journal article publications authored by their nursing associates. In 2024, Ascension launched a database index to highlight its nurse authors, bring attention to subject matter expertise, foster collaboration among authors, and recognize impact within the profession. The work was the result of a collaborative effort among Ascension librarians and nurse research scientists.

The index uses the open-access software platform Zotero, intended for reference management (Corporation for Digital Scholarship, 2024), with a public-facing cloud option to minimize expenses. This unconventional use of the platform capitalizes on the software's bibliographic database management capabilities while allowing for institutional-specific metadata for internal reporting. To populate the index, publications are captured through two primary methods. A REDCap survey is widely shared with nurses for self-reporting and a PubMed search alert is used to capture publications proactively. All identified publications are subsequently verified before inclusion. Eligible publications are easily added to the index. The platform's browser extension is used to capture bibliographic information from open-access websites, typically PubMed, with a single click. This process minimizes manual data entry, making it possible for a single librarian to manage the work of adding new publications to the index on a monthly basis.

The Ascension Nurse Author Index is making an impact across Ascension's multi-state healthcare system inclusive of approximately 48,000 nurses:

- Nurse authors are promoted by highlighting the index's availability through internal communications during Nurses Week.
- An executive-authored letter of appreciation is sent to all associate nurses authoring a newly published article included in the index.
- "Author's Row," a feature on the system-level nursing intranet site relies on the index to identify and highlight recently published nurse authors.
- Annual and quarterly reports use the index to quantify nurse authorship within the organization.
- The value of nurse scholarship is publicized across the healthcare system.

By creative use of the open access platform, clinical librarians can successfully partner to create value for their organization by highlighting the work of its nurses. The nursing shortage (AACN, 2023; National Center for Health Workforce Analysis, 2022) combined with the

DOI: dx.doi.org/10.5195/jmla.2025.2086

concern that many hospitals are reassessing the value of library services (Harrow et al., 2019), makes it an opportune time for librarians in the clinical environment to support and showcase the work of the nurse.

AUTHOR CONTRIBUTIONS

Aida Marissa Smith, MLIA, AHIP: Conceptualization, Project administration, Methodology, Writing – original draft, Writing – review & editing. Alexia Estabrook, MSLS: Conceptualization, Methodology, Writing – review & editing. Mary A. Hyde, MSLS, AHIP: Conceptualization, Methodology, Writing – review & editing. Michele Matucheski, MLIS, AHIP: Conceptualization, Methodology, Writing – review & editing. Eleanor Shanklin Truex, MLIS, RN: Conceptualization, Methodology, Writing – review & editing.

REFERENCES

1. AACN. (2023). *Nursing Workforce Fact Sheet*. https://www.aacnnursing.org/news-data/fact-sheets/nursing-workforce-fact-sheet.
2. Ascension. (2024). *The Ascension Nurse Author Index* [Dataset]. https://www.zotero.org/groups/5387431/ascension_nurse_author_index/library.
3. Corporation for Digital Scholarship. (2024). *Zotero* [Computer software].
4. Harrow, A., Marks, L. A., Schneider, D., Lyubechansky, A., Aaronson, E., Kysh, L., & Harrington, M. (2019). Hospital library closures and consolidations: A case series. *Journal of the Medical Library Association: JMLA*, *107*(2), 129–136. https://doi.org/10.5195/jmla.2019.520
5. National Center for Health Workforce Analysis. (2022). *HRSA Nurse Workforce Projections, 2020-2035*. Health Resources and Services Administration. https://bhw.hrsa.gov/sites/default/files/bureau-health-workforce/Nursing-Workforce-Projections-Factsheet.pdf.

AUTHORS' AFFILIATIONS

Aida Marissa Smith, MLIA, AHIP, aida.smith@ascension.org, https://orcid.org/0000-0002-5933-9848, Ascension Missouri, MO

Alexia Estabrook, MSLS, alexia.estabrook@ascension.org, Ascension Providence Hospital Michigan, MI

Mary A. Hyde, MSLS, AHIP, mary.hyde@ascension.org, Ascension Florida, FL

Michele Matucheski, MLIS, AHIP, michele.matucheski@ascension.org, Ascension Wisconsin, WI

Eleanor Shanklin Truex, MLIS, RN, eleanor.truex@ascension.org, https://orcid.org/0009-0000-6385-2516, Ascension Illinois, IL

Received October 2024; accepted October 2024

ISSN 1558-9439 (Online)

VIRTUAL PROJECT
DOI: dx.doi.org/10.5195/jmla.2025.2084

Use of large language model (LLM) to enhance content and structure of a school of dentistry LibGuide

Emily P. Jones

See end of article for authors' affiliations.

A librarian used a large language model (LLM) to revise a dentistry subject LibGuide. Prompts were used to identify methods for optimizing navigational structure for usability, highlight library-specific information students need additional help with, and write summaries of page content. Post-revision, LibGuide access increased, and students provided anecdotal feedback that they perceive the changes positively. LLMs may enhance LibGuide discoverability and usability without adding significant time and resource burdens for librarians.

Keywords: Generative AI; Artificial Intelligence (AI); LibGuides; Large Language Models

Virtual Projects are published on an annual basis in the *Journal of the Medical Library Association* (*JMLA*) following an annual call for virtual projects in *MLAConnect* and announcements to encourage submissions from all types of libraries. An advisory committee of recognized technology experts selects project entries based on their currency, innovation, and contribution to health sciences librarianship.

BACKGROUND

Large language models (LLMs) like Chat-GPT [1] and Claude.ai [2] are useful tools for summarizing, predicting, and generating text. These tools have potential to increase productivity and decrease the time burden of common, text-based tasks for librarians like LibGuide content creation.

VIRTUAL PROJECT DESCRIPTION

In June 2024, a librarian used an LLM, Claude.ai, to facilitate a major redesign of a dentistry LibGuide. Through a series of prompts, the librarian consulted the LLM to generate introductions summarizing content of specific pages and to restructure the LibGuide, formerly organized by resource format. Screenshots of the LibGuide pre- and post-revision, as well as examples of prompts provided to, and responses received from, Claude.ai are accessible via the author's institutional repository.

There was a 131% increase in LibGuide access from June - September 2024 (n = 2,288) compared to the same period the year before (n = 989). To the author's knowledge, no other changes were made that would significantly impact usage like new outreach or instruction. In addition to the increase in usage statistics, students have provided anecdotal feedback that they perceive the LibGuide to be more user-friendly and useful after the revision.

DISCUSSION

LLMs are cost-effective, as most are free, low-cost, or institutionally provided, and time-saving. Large amounts of text can be generated in a matter of seconds, whereas comparable output by a librarian may take hours. Additionally, LLMs can be used across various aspects of medical librarianship across any discipline and can be used to generate or clarify text about complex research topics like systematic reviews and data management.

While there are advantages to using LLMs like increasing efficiency and productivity, there are challenges as well. Concerns have been raised about accuracy of responses, privacy, and algorithm bias [3]. While LLMs are skilled at text-based tasks, they may not be able to adequately produce responses that require nuance, context, or complexity of thought. Therefore, it is best practice to review LLM responses for clarity and accuracy before using them. Additionally, LLM responses are highly dependent upon prompts received. Responses also change each time they're provided, even if the same prompt is provided, whether by the same or different individuals.

CONCLUSION

This project demonstrates a practical example of how librarians can apply generative artificial intelligence (AI) technologies to routine tasks like LibGuide revision and content creation. Using LLMs to develop page

DOI: dx.doi.org/10.5195/jmla.2025.2084

introductions and to reorganize content resulted in a usage increase. While not causative, while not causative, this increase may be correlated to increased discoverability and usability from using generative AI developed text and suggestions. LLMs can enhance the instructional component of LibGuides without adding a significant time burden for the creator.

REFERENCES

1. OpenAI. ChatGPT-4 [Internet]. 2024 [cited 2024 Sep 18]. Available from: https://chatgpt.com/.
2. Anthropic. Claude 3.5 Sonnet [Internet]. 2024 [cited 2024 Sep 18]. Available from: https://claude.ai/new.
3. Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng P-C, Bright TJ, Tatonetti N, Won KJ, Gonzalez-Hernandez G, Moore JH. ChatGPT and large language models in academia: opportunities and challenges. BioData Min. 2023 Jul 13;16(1):20.

AUTHORS' AFFILIATIONS

Emily P. Jones, epjones3@email.unc.edu, https://orcid.org/0000-0002-4294-7564, University of North Carolina at Chapel Hill (UNC) Health Sciences Library, Chapel Hill, NC

Received October 2024; accepted October 2024

ISSN 1558-9439 (Online)

VIRTUAL PROJECT
DOI: dx.doi.org/10.5195/jmla.2025.2078

Development of an open access systematic review instructional video series accessible through the SPI-Hub™ website

Sheila V. Kusnoor; Annette M. Williams; Taneya Y. Koonce; Poppy A. Krump; Lori A. Harding, AHIP; Jerry Zhao; John D. Clark; Nunzia B. Giuse, FMLA

See end of article for authors' affiliations.

Given the key role of systematic reviews in informing clinical decision making and guidelines, it is important for individuals to have equitable access to quality instructional materials on how to design, conduct, report, and evaluate systematic reviews. In response to this need, Vanderbilt University Medical Center's Center for Knowledge Management (CKM) created an open-access systematic review instructional video series. The educational content was created by experienced CKM information scientists, who worked together to adapt an internal training series that they had developed into a format that could be widely shared with the public. Brief videos, averaging 10 minutes in length, were created addressing essential concepts related to systematic reviews, including distinguishing between literature review types, understanding reasons for conducting a systematic review, designing a systematic review protocol, steps in conducting a systematic review, web-based tools to aid with the systematic review process, publishing a systematic review, and critically evaluating systematic reviews. Quiz questions were developed for each instructional video to allow learners to check their understanding of the material. The systematic review instructional video series launched on CKM's Scholarly Publishing Information Hub (SPI-Hub™) website in Fall 2023. From January through August 2024, there were 1,662 international accesses to the SPI-Hub™ systematic review website, representing 41 countries. Initial feedback, while primarily anecdotal, has been positive. By adapting its internal systematic review training into an online video series format suitable for asynchronous instruction, CKM has been able to widely disseminate its educational materials.

Keywords: Asynchronous learning; online learning; Systematic Reviews as Topic

Virtual Projects are published on an annual basis in the *Journal of the Medical Library Association* (*JMLA*) following an annual call for virtual projects in *MLAConnect* and announcements to encourage submissions from all types of libraries. An advisory committee of recognized technology experts selects project entries based on their currency, innovation, and contribution to health sciences librarianship.

CONTEXT, AIMS, AND SIGNIFICANCE OF PROJECT

In the Fall of 2023, the Center for Knowledge Management (CKM) at Vanderbilt University Medical Center (VUMC) released an online instructional video series on systematic reviews (https://spi-hub.app.vumc.org/sysrev/home) [1]. The project aimed to develop a freely available learning resource to ensure equitable access to quality educational materials on systematic reviews. The project further aimed to strengthen evidence-based medicine initiatives by equipping individuals with the knowledge and skills to design, conduct, and evaluate systematic reviews.

The idea for developing the online instructional video series emerged from an internal CKM training series on systematic reviews, which was conducted in Spring 2023. The training consisted of eight-hour-long sessions designed and led by three CKM team members experienced in conducting systematic reviews. The goal of the internal training was to strengthen the overall capacity of the CKM team to conduct and contribute to systematic reviews at VUMC. After the series ended, CKM team members discussed ways to share the training with other individuals and groups wanting to learn more about systematic reviews. The team ultimately chose to adapt the content developed for the internal training series into a video series format to allow for online, asynchronous instruction for an external audience.

The online systematic review training was created by CKM's team of information scientists and made available through the Scholarly Publishing Information Hub (SPI-Hub™); an open access tool developed by CKM. Although SPI-Hub™ was originally conceived with the intent of helping authors identify suitable journals in which to publish their work while steering them away from controversial journals, the site over time has become a

DOI: dx.doi.org/10.5195/jmla.2025.2078

truly comprehensive one-stop shop for researchers interested in learning more about the publishing industry, researcher profiles, upcoming trends on open access tools and policies, and now, through the link to the systematic reviews training, evidence provision. Additionally, as SPI-Hub™ consistently vaunts an average of 4,000 monthly national and international uses, its site was determined to be the perfect platform for the systematic review instructional videos [2].

PROJECT DESCRIPTION AND TECHNOLOGY

The SPI-Hub™ systematic review training website includes brief instructional videos covering all aspects of the systematic review process. Topics discussed include types of literature reviews, reasons for conducting a systematic review, how to develop a systematic review protocol, steps in conducting a systematic review, web-based tools to aid with the systematic review process, publishing a systematic review, and appraising systematic reviews.

In total, 18 videos, with an average length of 10 minutes each, were created over a period of approximately four months. The instructional videos are accompanied by quiz questions to allow users to check their understanding of the material. The questions were formatted to provide feedback and allow for unlimited re-attempts. Each video page also lists learning objectives and key references. The project team included five information scientists and two application developers; the initiative was led by the Center's director.

Content within the systematic review section of SPI-Hub™ is delivered through Laravel, an open-source PHP Framework, in conjunction with internally developed HTML5 scripts. The instructional videos were created using the PowerPoint screen recording feature.

ADVANTAGES, LIMITATIONS, AND IMPACT

By adapting CKM's internal systematic review training series into an online video series format, CKM has been able to reach a large and diverse audience. Usage data has been monitored to help assess the project's impact. From January through August 2024, there were 1,662 accesses to the systematic review instructional series, representing 41 countries. There was a notable peak in May, likely attributable to CKM's presentation at the Medical Library Association annual conference [3].

The CKM team has promoted the use of the SPI-Hub™ systematic review video series, including incorporating it in courses where appropriate. For example, the resource was shared with graduate students in a lecture on literature reviews for a biomedical informatics course on scientific communication. Additionally, the resource was shared with fourth-year medical students enrolled in the institution's WikiMed advanced elective.

While the SPI-Hub™ systematic review training website is fully implemented, the CKM team anticipates regularly assessing and responding to user feedback. Thus far, the feedback has been primarily anecdotal; however, all groups to whom the online systematic review instruction was presented have had an enthusiastic response, commenting on its comprehensiveness, ease of use, and the fact that it was freely available through the web.

AUTHOR CONTRIBUTIONS

Sheila Kusnoor: project leader, methodology, writing – original draft, writing – review & editing. Annette Williams: methodology, writing – original draft, writing – review & editing. Taneya Koonce: methodology, writing – original draft, writing – review & editing. Poppy Krump: methodology. Lori Harding: methodology. Jerry Zhao: methodology, software. John Clark: methodology, software. Nunzia Giuse: conceptualization, writing - original draft, writing – review & editing.

REFERENCES

1. Kusnoor SV, Williams AM, Koonce TY, Krump P, Harding L, Zhao J, Giuse NB. Systematic Reviews: A Comprehensive Overview Video Series. Vanderbilt University Medical Center, Center for Knowledge Management, 2023. Available from: https://spi-hub.app.vumc.org/sysrev/home. [cited 26 Sep 2024].
2. Koonce TY, Blasingame MN, Zhao J, Williams AM, Su J, DesAutels SJ, Giuse DA, Clark JD, Fox ZE, Giuse NB. SPI-Hub™: a gateway to scholarly publishing information. J Med Libr Assoc. 2020 Apr;108(2):286-294. DOI:10.5195/jmla.2020.815. Epub 2020 Apr 1. PMID: 32256240; PMCID: PMC7069808.
3. Kusnoor S, Williams A, Koonce T, Krump P, Harding L, Zhao J, Giuse N. All About Systematic Reviews: Making Quality Learning Equitable and Freely Accessible to All. Portland, OR: Medical Library Association Meeting, May 20, 2024.

AUTHORS' AFFILIATIONS

Sheila V. Kusnoor, PhD, sheila.v.kusnoor@vumc.org, https://orcid.org/0000-0001-6025-6057, Senior Information Scientist and Associate Director for Research, Center for Knowledge Management, Vanderbilt University Medical Center; Research Assistant Professor, Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

Annette M. Williams, MLS, Annette.williams@vumc.org, https://orcid.org/0000-0002-2526-3857, Senior Information Scientist and Associate Director for Metadata Management, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Taneya Y. Koonce, MSLS, MPH, taneya.koonce@vumc.org, https://orcid.org/0000-0002-4014-467X, Deputy Director, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

DOI: dx.doi.org/10.5195/jmla.2025.2078

Poppy A. Krump, MSIS, poppy.krump@vumc.org, https://orcid.org/0000-0002-3081-6487, Information Scientist, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Lori A. Harding, MLS, AHIP, la.harding@vumc.org, https://orcid.org/0000-0001-5241-2849, Information Scientist, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Jerry Zhao, MS, MLIS, jerry.zhao@vumc.org, Senior Application Developer, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

John D. Clark, MS, john.clark@vumc.org, Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Nunzia B. Giuse, MD, MLS, FACMI, FMLA, nunzia.giuse@vumc.org, https://orcid.org/0000-0002-7644-9803, Vice President for Knowledge Management, Director, Center for Knowledge Management, Vanderbilt University Medical Center; Professor of Biomedical Informatics and Medicine, Department of Biomedical Informatics, Vanderbilt University Medical Center; Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

Received September 2024; accepted October 2024

ISSN 1558-9439 (Online)

DOI: dx.doi.org/10.5195/jmla.2025.2030

Early explorations of holistic review in graduate medical education

Gena C. Dunivan; Jonathan D. Eldredge, AHIP, FMLA; Marlene P. Ballejos; Melissa Gonzales; Valerie Romero-Leggott

See end of article for authors' affiliations.

Background: Graduate Medical Education programs have implemented holistic review to improve the selection process for new residents. Holistic review will have a profound effect on Health Information Professionals (HIPs) with the arrival of medical residents with different backgrounds and needs. The unique experiences and skills of HIPs will position them well for the new realities in medical residency programs. This article traces the historic roots of holistic review.

Methods: The authors employed a scoping review to track the historical traces of holistic review in Graduate Medical Education over the formative period of 1999-2019.

Results: Medical residency programs over a 20-year period piloted holistic review in the screening, interview, and multiple time periods in the selection process. These ventures reflected a diversity of approaches and creative adaptations from other disciplines such as personnel management, organizational psychology, and active learning forms of education

Conclusion: Health information professionals and medical educators will better engage with the newer cohorts of residents when equipped with a history of holistic review.

Keywords: Holistic review; Graduate medical education; Internship and residency; School admission criteria; Health care disparities; Cultural diversity; Population groups; Personnel selection; Social determinants of health; Social justice

See end of article for supplemental content.

BACKGROUND

The practice of holistic review has been implemented in medical education in recent years, particularly in Graduate Medical Education (GME). Holistic review already has had a major impact on selection processes for new medical residents. Holistic review will continue to affect the composition of medical residency programs and will profoundly affect how medical educators and Health Information Professionals (HIPs) interact with these residents. HIPs (health sciences librarians, informaticists, informationists, and archivists) have a long-standing history of working closely with GME programs in the US. Historically, HIPs have conducted literature searches to support patient care, clinical research, or bedside instruction such as patient rounds in support of GME programs. HIPs also have ensured that authoritative information resources are available for GME faculty residents, and fellows [1]. The more informationist and informatics-oriented HIPs have evaluated point-of-care resources and electronic health records. Some of these same colleagues administer GME Clinical Informatics Fellowships. HIPs' management skills have led to their involvement in the selection, evaluation, and education oversight GME committees at their institutions [2]. Some HIPs are sought out by GME programs for their curricular and instructional design skills. Over the past decade, HIPs have been closely connected with teaching medical residents and fellows in accordance with specialty-based ACGME Milestones [3]. HIPs also have demonstrated a commitment to diversity, equity, and inclusion for at least half a century [4-7]. These factors converge to make a deeper understanding of holistic review highly relevant to HIPs and their medical educator colleagues.

This scoping review traces the historical antecedents of current holistic review practices. It explains how these antecedents led to the quick and, perhaps surprising, rapid acceptance of holistic review.

The phrase "holistic review" in medical education refers to selecting candidates who will be well-matched to both the training program and to the program's patient populations. The Association of American Medical Colleges (AAMC) formally defines holistic review as a "flexible, mission-driven approach to recruit and assess an

DOI: dx.doi.org/10.5195/jmla.2025.2030

individual's competencies by considering their experiences, attributes, and metrics in order to select applicants who will best contribute to the program's unique goals, learning environment, and the practice of medicine [8]." The Accreditation Council for Graduate Medical Education (ACGME) in the United States (US) thoroughly embraced holistic review in late 2019 and encouraged Graduate Medical Education (GME) programs nationwide to institute holistic review in selecting applicants for specialty medical training. Societal concerns about health disparities fueled by health inequities disproportionately affecting women and minorities largely prompted and accelerated ACGME's commitment to holistic review [9].

Examining these early explorations of holistic review within GME programs holds significant historical and practical value to everyone connected to medical education in the US. In practical terms, the great enthusiasm for holistic review has not yet translated to many peer reviewed research articles on the subject so these early explorations have added pragmatic value. HIPs and clinical educators can learn a great deal about the early attempts to employ holistic review as they all probably have (or soon will be) been called upon to institute these practices given the groundswell in interest among GME programs. Finally, innovators in medical education can learn from how these, at the time, radical departures from standard selection practices contributed to a far-reaching reform movement in GME. In short, readers can gain valuable insights from this recent history.

Most readers associated with medical education recognize the profound influence of the residency Match process upon medical school curricula and medical students in the US. The Match was invented in 1951 to prevent medical students from exploitation by medical residencies. At that time, there were twice as many medical school graduates as there were available slots in residency programs, causing a severe power imbalance against the medical students. The Match consists of medical students submitting their lists of preferred ranked residency program choices while the medical residencies submit their preferred ranked choices of graduating medical students [10-12].

While this complementary ranking system stood the test of time, by the early 2000s the Match unintentionally led to mismatches of applicants with selected medical residencies. Part of the problem was the large number of residency programs to which individual medical students would apply. Faced with so many applicants, residency program faculty sought more efficient ways to streamline their screening processes. A quick yet unfortunate metric chosen to screen prospective residents turned out to be the United States Medical Licensure Exam (USMLE) Step 1. Residency programs in the process inadvertently succumbed to the quantitative fallacy by elevating an easily-measured score while ignoring other important factors [13-14]. The USMLE Step 1 has a mixed record in predicting students' success in residency programs [15]. Meanwhile, this one exam has caused unwarranted psychological stress for medical students [16]. As Salari and Deng note, "it is so symbolic that medical students fear the exam from the start of medical school." They continue, "the infamous nature of Step 1 originates from the substantial weight scores carry in the resident selection process (p. 1,312)." rather than as an intended "checkpoint" of student progress in acquiring knowledge of the basic sciences [17]. In 2014, a total of 94% of medical residency program directors considered Step 1 to be a major factor in screening applicants. This overwhelming emphasis had the unintended consequence of suppressing needed medical school curricular change [18]. Importantly, the USMLE Step 1 perpetuated racial and gender disparities [19].

Diversity is of particular importance across all specialties as research supports that the quality of health care is improved when the providers' backgrounds or ethnicities reflect the community they serve [20-21]. Health care is far from this target: although minority groups represent 33% of the overall US population, the physician workforce is comprised of only 4.1% Black, 4.4% Hispanic, and 0.4% American Indian/Alaskan Native[22].

Fortunately, the USMLE Step 1 transitioned in 2022 to a pass/fail rather than scored exam. Holistic review emerged as a novel opportunity to align acceptances of medical school graduates with the needs of the residency programs and the needs of the patients served by these programs.

It will take a sustained, long-term effort to dismantle the structural and systemic sources that perpetuate subtle forms of racism in the US. The roots of racism can be traced back to the Doctrine of Discovery in the 1400s in Spain and Portugal then later adopted by other European nations and in the American colonies [23-25]. These structures and systems have taken centuries to accumulate and become entrenched [26]. Holistic review promises to erode these systems and to facilitate better matches between the needs of residency programs and their patients. In so doing, medical residencies will serve society's evolving needs [27].

Research Question

What insights can be gained from early explorations into holistic review practices for selecting new residents in GME residency programs in the US prior to the broad endorsement of this approach in late 2019?

METHODS

The authors selected the scoping review methodology to address the research question by providing a rapid, "preliminary map of the literature [28]." to understand the

DOI: dx.doi.org/10.5195/jmla.2025.2030

evolution of the revolutionary practice of holistic review. The authors referred to the Scoping Reviews protocol extensions to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA-ScR) [29] and the Joanna Briggs Institute *Manual for Evidence Synthesis* throughout this scoping review [30].

Data Sources and Searching

This scoping review presented challenges requiring creativity to retrieve the relevant early explorations of holistic review in GME. A search of PubMed on May 30, 2024 illustrates the challenges. The combined MeSH and adjacent textword search strategy of "Education, Medical, Graduate"[Majr] AND "holistic review" produced 68 references in PubMed. Only five references were published prior to 2020, although the concept of holistic review had existed for about two decades. The first use of the phrase "holistic review" among these five references that matched the formal concept first appeared in a 2016 article [31].

Three databases were searched to locate relevant studies due to their coverage of medical education references: Ebsco Cumulative Index to Nursing and Allied Health Literature (CINAHL), Ebsco Education Research Complete, and PubMed from the National Library of Medicine. The PRISMA flow diagram for scoping reviews in **Figure 1** and the detailed search strategies in the online **Appendix** provide many details.

All of the final searches in the three databases employed combinations of controlled vocabularies, keyword, and truncated keyword approaches. The first set of searches retrieved references intended to capture the concept of graduate medical education. The second set of searches identified references connected with either academic admissions or personnel selection since elements of both can occur within holistic review involving prospective residents. The final set of searches captured the range of attributes associated with holistic review. The three searches conducted during August 2020 were combined with AND to produce 635 references that were then filtered by the publication years 1999-2019 resulting in 513 references. The start date of 1999 was chosen because it was the first known attempt at holistic review and the end date of December 2019 was selected due to the endorsement of holistic review by ACGME [32]. This methodology text below and the **Appendix** will allow others to reproduce these searches in the future, per PRISMA guidelines and checklists for scoping reviews [33-34]. The Peer Review of Electronic Search Strategies (PRESS) checklist [35] proved helpful for replicability purposes even though it was originally intended for systematic reviews. The 513 references were uploaded into Rayyan, a web application for reviewing and critically appraising references. It should be noted for future searchers wishing to replicate this scoping review, that PubMed eclipsed the other two databases that either duplicated titles or contained articles to be excluded from final consideration. Supplementary tables in the online **Appendix** detail the initial reasons for rejecting the vast majority of the initially retrieved references.

Inclusion and Exclusion Criteria

The first two authors (GD and JE) reviewed the abstracts of the 513 references in Rayyan with their choices concealed from one another. The authors rejected the majority of references (n = 444) mainly because they involved populations other than prospective medical residents (undergraduate medical education, dental, nursing, allied health, post-residency fellows, etc.), were from outside the US, or were preliminary works exploring the possibilities of holistic review; they also excluded editorials or commentaries. They agreed on all but 30 references when working in isolation with their choices concealed from one another. They reviewed these 30 references in Rayyan together in a meeting and resolved any differences. They agreed on 69 references to advance to the next stage. They attached PDFs of the actual articles to these 69 references for the next review phase.

Later, the authors removed 40 references on further examination of the articles themselves as these pertained more to background or rationales for holistic review.

RESULTS

The remaining 19 selected articles within this scoping review reflect the wide range of diverse holistic review practices involving prospective medical residents during 1999-2019 in the US. **Table 1** outlines the historical progression of explorations with holistic review. The earliest instance of a holistic review was published by Thomas in 1999. It sought to correct for minorities and women historically underrepresented in medical residencies [36]. In the later periods, there were many more examples to consider.

Holistic review has been utilized at several time points: screening, the interview, and at multiple time points.

Table 1 Chronology of Holistic Review Efforts

Year	First Author	Summary
1999	Thomas	Orthopedics department leadership insisted on greater gender and racial representation.
2009	Quintero	Diversified pool of orthopedic residents by selecting a range of Myers-Briggs personality types.
2010	Hemaida and Kalb	Decided on new residents based on non-cognitive factors and

DOI: dx.doi.org/10.5195/jmla.2025.2030

		interpersonal skills identified during interviews.
2011	Bell et al.	Matched prospective residents' personalities to the surgery department's composite profile.
2015	Stephenson-Famy	Attributes of surgery residency interviews most likely to lead to successful residents.
2016	Schenker et al.	Validated a standard interviewing protocol that provided a more comprehensive picture of orthopedic surgery candidates.
2017	Bowe et al.	Family practice residency program directors identified traits of successful residents then used to create a ranking form to evaluate candidates.
2017	Martin and Salzberg	Delphi study to determine most desirable traits in family medicine residents.
2018	Schnapp	Emergency medicine candidates recorded a three-minute video designed to identify interpersonal skills and professionalism.
2019	McGuire	Structured interviews to identify desired traits and motivators for possible orthopedics residents.
2019	Shebrain	Compared cognitive with non-cognitive aspects of surgery residence candidates. Determined that cognitive traits were more important.
2019	Albana et al.	Increased diversity in internal medicine department by debiasing the interviewers and elevating desired non-academic traits in candidates.
2019	Butler et al.	Greater emphasis upon urology candidates' attributes and experiences and less emphasis upon academic performance.
2019	Garrick et al.	Eliminated USMLE Step One cutoff scores, increased diversity in interviewers, and recruited more diverse emergency medicine candidates.
2019	Spottswood et al.	Created comprehensive diversity recruitment plan that included a pipeline program, faculty diversity, and a standard selection process to recruit desirable radiology candidates.
2019	Wusu et al.	Structured interviews in family medicine residency did not include candidates' academic records and created a climate comfortable to candidates from underrepresented backgrounds.
2019	Byrd	Standardized video interviews to assess competencies in interpersonal communication and professionalism in emergency medicine candidates.
2019	Spector	Redesigned the interview process to remove potential bias in a neurology residency program.
2019	Villwock	Informatics techniques to evaluate emotional intelligence and personality profiles for a otolaryngology residency program.

Screening

The strategies utilized for the screening process focused on standardization of the process and options such as Standardized Video Interviews (SVI). Bird et al utilized the Association of American Medical Colleges (AAMC) SVI to help screen candidates for interviews by focusing on competencies such as interpersonal communication skills and professionalism, allowing the selection committee to differentiate candidates effectively [37]. Spector and Railey described their process to improve representation of under-represented in medicine (URiM) candidates, by implementing a process with no USMLE cut score. They used a point system for evaluating applicants for other characteristics based on other factors such as extracurricular activities/leadership, letters of recommendations, and life experiences. They reported the racial/ethnic discrepancies in interview offers decreased from 10.6% to 3.6% due these efforts [38]. Finally, Villwock et al. described the Selection Tool for Applicants to Residency (STAR) that uses a predetermined criteria algorithm to score different aspects of the applicants' Electronic Residency Application Service (ERAS) submissions to create an initial ranking for all applications. STAR weighs candidate characteristics on a 0 to10 point scale with weightings in parenthesis() : academic (1x), extracurricular activities (1x), research experience (3x), leadership positions (3x), and geographic connection (1x) to meet program needs. STAR offers a potentially more efficient selection process that avoids missing "diamonds in the rough" candidates who otherwise could be missed by focusing too intensely on USMLE step scores. Importantly, STAR did not disqualify eligible URiM and women candidates invited for an interview and there was no difference in resident attrition rates [39].

Interview

Holistic review methods at the interview stage of selecting candidates for residencies rely on a range of approaches largely from the domains of personnel management and psychometric tests. Most of these holistic interviewing

DOI: dx.doi.org/10.5195/jmla.2025.2030

methods try to overcome interviewer subjectivity by providing structured formats. The interviewing methods described generally are future-oriented toward residents' actual specialty practice responsibilities rather than relying on past performance on standardized exams.

Bell et al. employed the TriMetrix Personal Talent Report (TPTR) to create an inventory of a surgery department's "behavioral styles, intrinsic motivators, and dimensional balance Page 534)" and a list of characteristics of superior performance. The characteristics in this profile that emerged pointed to the inadequacy of grades, exam scores, letters, and other traditional selection methods to predict success in residency (Pages 536 and 539) [40].

Bowe et al. created an Applicant Ranking Tool (ART) that aligned traits with the six ACGME competency areas along with essential non-cognitive areas gleaned from the interview including conscientiousness, curiosity, interpersonal skills, confidence, and recognition of one's limits. Additionally, creating an ART for one specialty does not necessarily mean it will translate to other specialties due to the different knowledge, skills, and values emphasized within varied specialties [41].

Hemaida and Kalb applied the Analytic Hierarchy Process (AHP), which offers an aid to making complex decisions, in selecting Family Practice residents. Family practice faculty, residents, and administrators who interacted with residents were asked to identify the most important six factors from a list of activities connected with selecting new residents. These factors were interview assessments, interpersonal skills, fit with current team, conformity with the organizational culture of the program, their personal statement, and alignment of future practice plans with program areas of emphasis. Respondents did not rank highly the prospective residents' medical school grades or USMLE scores [42]. Forman has provided an excellent background description of the AHP [43].

McGuire et al. structured prospective resident interviews according to critical tasks based upon a job analysis of successful orthopedic surgeons. The structured interviews also sought to determine the prospective residents' capabilities and motivations [44].

Quintero et al. used a prospective cohort study to identify bias in the selection of orthopedic surgeons by comparing the Myers-Briggs personality types of interviewers and interviewees. This can be employed to avoid too much convergence of personality types within a program [45]. Past studies have noted the clustering of certain personality profiles within any given medical specialty [46-48]. Schnapp et al. used a Standardized Video Interview (SVI) to evaluate 125 prospective emergency medicine residents' interpersonal and professionalism skills. Candidates video recorded their three-minute responses to a series of six standard questions posed by program faculty members. These scores were compared with faculty gestalt scores based upon in-person unstructured faculty interviews. There was no significant correlation between the SVI scores and the faculty gestalt scores, suggesting that the two formats are measuring different aspects of professionalism and interpersonal skills [49].

Shebrain et al. discovered that USMLE Steps 1 and 2 scores had an inverse relationship in their program for predicting candidates' success during in-person interviews. Non-cognitive aspects, particularly those presented during the interview were far more predictive of resident candidate success in securing favorable rankings in the interview scores. The non-cognitive aspects consisted of letters of recommendation, personal statements, how the candidate represented oneself, the candidate's stated interest in the specific program, responses to standardized questions, and the degree of connection an interviewer felt with the interviewee. This latter connection proved to be the only predictive factor when controlling for all other variables [50].

Stephenson-Famy et al. determined the importance of the interview in the selection process based on a literature review of 104 studies. The interview assesses non-cognitive attributes such as communications skills, maturity, and professionalism. The authors recommended that a rigorous and structured interview strategy should replace the unblinded and unstructured interviews of the past. The structured interview should include: a written description of desired traits, standardized questions, behavior-specific anchors with a scoring rubric, multiple observers, interview trainings to avoid unethical questions, and blinding of the interview to academic metrics [51].

Johns Hopkins Department of Orthopedic Surgery committed to a holistic review of prospective candidates that emphasized candidates' potential to succeed rather than their national exam score metrics. Candidates' diverse backgrounds, interpersonal skills, openness to new learning, and work ethic instead were the focus of onsite interviews. African-American and women residents scored no differently than other residents in multiple assessment events and the board certification exams [52].

Multiple Time Points

Many of the studies that utilized a holistic process included similar key features in a multipronged approach at different times during the application process, including recruitment, screening, interviewing, and ranking. Outreach to the desired applicant pools (For example, URiM medical students) was mentioned by several articles [53-54]. Having standardized "themes" to discuss in interviews was discussed by Schenker et al. Each interviewer was assigned a theme including knowledge, affective domain, ethics, research, and "fit" [55]. Most articles also emphasized the importance of standardized screening tools that decreased emphasis on

DOI: dx.doi.org/10.5195/jmla.2025.2030

traditional academic metrics coupled with structured interviews [56]. A unique component discussed in these articles was that the representation of URiM residents and faculty in the process was critical. This could include being present at recruitment events, interviews or sending a personal email or follow-up phone call [57]. Garrick et al. highlighted an annual diversity recruitment dinner in which URiM applicants in emergency medicine, internal medicine, and surgery could attend to demonstrate the hospital's support for diverse residents [58]. Also discussed was sponsoring a no cost "second-look weekend" for highly desired applicants [59]. These articles exemplify a multi-pronged approach to holistic review while working to increase diversity in a program. An example highlighting this success can be seen in Garrick et al. Not only did their multi-pronged approach increase the proportion of URM graduating from their residency (12% to 27%) over an eleven-year period, but the authors note that during that time all residents graduated on time and the program has a first time pass rate of 98%.

LIMITATIONS

There are a number of limitations to this study. First, authors of these documented studies varied in their operational definitions of "holistic review" practices. These authors generally referred to practices intended to review prospective medical residents beyond primarily exam scores or other systems rigidly adhering to quantitative scores or comparative ranking as holistic review. Second, this scoping review included only studies conducted prior to the Covid-19 pandemic, thereby excluding all of the subsequent changes in medical residency recruitment, screening, and selection practices. Third, the identified study designs and approaches of holistic review varied widely, posing challenges for any close comparisons. In addition, the limited numbers of studies make it difficult to draw meaningful outcomes.

In addition, when programs are seeking to improve the number of URiM candidates that apply and ultimately match with their programs, other tools identified to increase success, included visiting clerkships, directed outreach to national organizations, voluntary presence of faculty and/or residents that identify as URiM at interviews with personal follow-up, and expense paid second look visits.

Despite the lack of standardization in how residency programs approach the concept of holistic review, it is clear that new processes are needed, especially now that the USMLE Step 1 has transition to pass/fail. The AAMC provides tools and resources for programs to begin crafting their unique approach to holistic review. It will be critical to follow outcomes to determine if the goals of holistic review have been achieved, including training residents who will thrive in their particular specialty, reflect the communities they serve, and broaden diversity across training programs to improve health access and outcomes for diverse populations.

CONCLUSION

Holistic review should be implemented throughout the recruitment, screening, interviewing, and ranking phases. Utilizing it at only one time point may still result in bias and hinder some candidates from moving to the next stage of selection. HIPs often serve on these selection committees and can offer their own social justice perspectives. Holistic review offers concrete ways to further diversify their ranks through recruitment and hiring efforts.

The main themes and recommendations include training and buy-in from faculty and other stakeholders involved in the recruitment efforts, development of qualities desired from candidates, attributes felt to be associated with success in the field, mission alignment, and priorities for individual programs such as commitment to an underserved area or bilingual proficiency. Standardized screening based on a predetermined scoring system with a decreased focus on academic metrics, coupled with structured interviews that are masked to academic metrics should be considered along with a decision aid or ranking tool that allows differential weights to be applied to particular capabilities that the program deems as "highly important". Finally, it is critical to follow outcomes to determine if the goals of holistic review have been achieved, including training residents that will thrive in their particular specialty, reflect the communities they serve, and broaden diversity across training programs to improve health access and outcomes for diverse populations.

Given the variability in the literature, it is somewhat difficult to offer an integrated synopsis. Key take-aways are that the majority of the 1999-2019 literature focuses on "fit" into residency and attempting to predict who will be successful in that residency. Performing holistic review during the screening process mainly highlighted ways that programs have decreased reliance on traditional academic metrics which did result in increased diversity in the applicants offered interviews. The main goal of utilizing a holistic process during the interview phase was to implement strategies to reduce interviewer subjectivity and unconscious bias. These strategies included standardized questions, bias training, and a variety of ranking tools. Holistic review involving multiple time points typically utilized standardized tools. All of these studies point to the need for longer term follow-up to both assess if utilization of holistic review tools resulted in more inclusive applicants and residents, but particularly if holistic review helps identify people who will ultimately be successful in that field. With their pragmatic sensibilities and their long-standing history of a commitment to equity and inclusion, HIPs can

DOI: dx.doi.org/10.5195/jmla.2025.2030

immediately grasp the benefits of and pivot to collaborate with the new residents selected due to holistic review.

AUTHOR CONTRIBUTIONS

Gena C. Dunivan, MD: Gena Dunivan contributed to the initial conceptualization of this project, designed and implemented the scoping review methodology, managed the search results in the Rayyan screening software, analyzed the screened results, created the search results tables, managed and coordinated as well as conducted oversight and leadership for this project, and wrote and then edited successive drafts of the manuscript.

Jonathan D. Eldredge (corresponding author): Jon Eldredge contributed to the initial conceptualization of this project, designed and implemented the scoping review methodology, conducted the searches, managed the search results in the Rayyan screening software, analyzed the screened results, created the search results tables 1-5, designed the PRISMA flowchart, and wrote then edited successive drafts of the manuscript.

Marlene P. Ballejos, PhD: Marlene Ballejos contributed to the preparation, creation, and presentation of the published work, particularly through critical review, commentary, and revisions during both pre- and post-publication stages.

Melissa Gonzales, PhD, MS: Melissa Gonzales contributed to the initial conceptualization of this project, designed the scoping review methodology, analyzed the screened results, created the search results tables, and wrote then edited the initial draft of the manuscript.

Valerie Romero Leggott: VRL contributed to the conceptualization of the initial work and critical review, commentary, and revisions, including Tables which contributed to the preparation and presentation of the published work

REFERENCES

1. Jones DA. Management of and issues specific to hospital libraries. In: Introduction to health sciences librarianship. Wood MS, ed. New York, NY: The Haworth Press, 2008: 341-67.
2. Henderson ME. New roles and new horizons for health sciences librarians and libraries. In: Health sciences librarianship. Wood MS, ed. Chicago, IL: Medical Library Association and Lanham, D: Rowman & Littlefield, 2014: 403-18.
3. Rhue DJ, Eldredge JD. Medical Residency Milestones: Competencies in Informatics, Library, and Evidence-Based Practice. *Med Ref Serv Q*. 2022;41(3):236-247. DOI:10.1080/02763869.2022.2093545
4. Samek T. *Intellectual Freedom and Social Responsibility in American Librarianship, 1967-1974*. McFarland; 2001.
5. Freeman RS, Hovde DM. *Libraries to the People: Histories of Outreach*. McFarland & Co.; 2003.
6. Baker D, Evans W. Libraries, society, and social responsibility. In: *Libraries and society: role, responsibility and future in an age of change*. DP Baker, ed. New York, NY: Elsevier Science & Technology. Accessed July 6, 2024. http://site.ebrary.com/id/10816534.
7. Schuessler J. Helping black culture rise upon a mountain of books. New York Times. Weekend Arts. 2024 Jul 7: C1 & C8-C9.
8. Association of American Medical Colleges. Holistic Review. Available from: <https://www.aamc.org/services/member-capacity-building/holistic-review >. Accessed 30 May 2024.
9. McDade WA. Increasing Graduate Medical Education Diversity and Inclusion. *J Grad Med Educ*. 2019;11(6):736-738. DOI:10.4300/JGME-D-19-00760.1
10. Chae SH. Is the match illegal? *N Engl J Med*. 2003;348(4):352-356. DOI:10.1056/NEJMsb021241
11. Wright AA, Katz IT. Match point? *N Engl J Med*. 2004;351(12):1165-1167. DOI:10.1056/NEJMp048122
12. Hendren WH. Is the match illegal? *N Engl J Med*. 2003;348(22):2259-2262.
13. Carmody JB. On Residency Selection and the Quantitative Fallacy. *J Grad Med Educ*. 2019;11(4):420-421. DOI:10.4300/JGME-D-19-00453.1
14. O'Mahony S. Medicine and the McNamara fallacy. *J R Coll Physicians Edinb*. 2017;47(3):281-287. DOI:10.4997/JRCPE.2017.315
15. Lombardi CV, Chidiac NT, Record BC, Laukka JJ. USMLE step 1 and step 2 CK as indicators of resident performance. *BMC Med Educ*. 2023;23(1):543. Published 2023 Jul 31. DOI:10.1186/s12909-023-04530-8
16. Moynahan KF. The Current Use of United States Medical Licensing Examination Step 1 Scores: Holistic Admissions and Student Well-Being Are in the Balance. *Acad Med*. 2018;93(7):963-965. DOI:10.1097/ACM.0000000000002101
17. Salari S, Deng F. A Stepping Stone Toward Necessary Change: How the New USMLE Step 1 Scoring System Could Affect the Residency Application Process. *Acad Med*. 2020;95(9):1312-1314. DOI:10.1097/ACM.0000000000003501
18. Gliatto P, Leitman IM, Muller D. Scylla and Charybdis: The MCAT, USMLE, and Degrees of Freedom in Undergraduate Medical Education. *Acad Med*. 2016;91(11):1498-1500. DOI:10.1097/ACM.0000000000001247
19. Rubright JD, Jodoin M, Barone MA. Examining Demographics, Prior Academic Performance, and United States Medical Licensing Examination Scores. *Acad Med*. 2019;94(3):364-370. DOI:10.1097/ACM.0000000000002366
20. Wusu MH, Tepperberg S, Weinberg JM, Saper RB. Matching Our Mission: A Strategic Plan to Create a Diverse Family Medicine Residency. *Fam Med*. 2019;51(1):31-36. DOI:10.22454/FamMed.2019.955445
21. Boylan M, Grant RE. Diversity and professional excellence. *J Natl Med Assoc*. 2004;96(10):1354-1362.

DOI: dx.doi.org/10.5195/jmla.2025.2030

22. Garrick JF, Perez B, Anaebere TC, Craine P, Lyons C, Lee T. The Diversity Snowball Effect: The Quest to Increase Diversity in Emergency Medicine: A Case Study of Highland's Emergency Medicine Residency Program. *Ann Emerg Med*. 2019;73(6):639-647. DOI:10.1016/j.annemergmed.2019.01.039

23. Jones RP. *The Hidden Roots of White Supremacy: And the Path to a Shared American Future*. First Simon & Schuster hardcover edition. Simon & Schuster; 2023.

24. Kendi IX. *How to Be an Antiracist*. One World; 2019.

25. Came H, Griffith D. Tackling racism as a "wicked" public health problem: Enabling allies in anti-racism praxis. *Soc Sci Med*. 2018;199:181-188. DOI:10.1016/j.socscimed.2017.03.028

26. Butler PD, Fowler JC, Meer E, Rosen IM, Reyes IM, Berns JS. A Blueprint for Increasing Ethnic and Racial Diversity in U.S. Residency Training Programs. *Acad Med*. 2022;97(11):1632-1636. DOI:10.1097/ACM.0000000000004847

27. Ludmerer KM. *Let Me Heal: The Opportunity to Preserve Excellence in American Medicine*. Oxford University Press; 2015. Accessed May 31, 2024. http://site.ebrary.com/id/10901481

28. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *International journal of social research methodology*. 2005;8(1):19-32. DOI:10.1080/1364557032000119616 [Page 21].

29. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med*. 2018;169(7):467-473. DOI:10.7326/M18-0850.

30. Aromataris E, Munn Z, Eds. JBI Manual for Evidence Synthesis. JBI, 2020. Available from https://synthesismanual.jbi.global. https://doi.org/10.46658/JBIMES-20-01 Accessed 4 January 2022.

31. Love JN, Ronan-Bentle SE, Lane DR, Hegarty CB. The Standardized Letter of Evaluation for Postgraduate Training: A Concept Whose Time Has Come? *Acad Med*. 2016;91(11):1480-1482. DOI:10.1097/ACM.0000000000001352

32. McDade WA. 2019. Op Cit.

33. McGowan J, Straus S, Moher D, et al. Reporting scoping reviews-PRISMA ScR extension. *J Clin Epidemiol*. 2020;123:177-179. DOI:10.1016/j.jclinepi.2020.03.016

34. Equator Network. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Available from: < https://www.equator-network.org/reporting-guidelines/prisma-scr/ >. Accessed 30 August 2024.

35. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *J Clin Epidemiol*. 2016;75:40-46. DOI:10.1016/j.jclinepi.2016.01.021

36. Thomas CL. African Americans and women in orthopaedic residency. The Johns Hopkins experience. *Clin Orthop Relat Res*. 1999;(362):65-71.

37. Bird SB, Hern HG, Blomkalns A, Deiorio NM, Haywood Y, Hiller KM, Dunleavy D, Dowd K. Innovation in Residency Selection: The AAMC Standardized Video Interview. Acad Med. 2019 Oct;94(10):1489-1497. DOI: 10.1097/ACM.0000000000002705. PMID: 30870151.

38. Spector AR, Railey KM. Reducing Reliance on Test Scores Reduces Racial Bias in Neurology Residency Recruitment. J Natl Med Assoc. 2019 Oct;111(5):471-474. DOI: 10.1016/j.jnma.2019.03.004. Epub 2019 Mar 29.

39. Villwock JA, Hamill CS, Sale KA, Sykes KJ. Beyond the USMLE: The STAR Algorithm for Initial Residency Applicant Screening and Interview Selection. J Surg Res. 2019 Mar;235:447-452. DOI: 10.1016/j.jss.2018.07.057.

40. Bell RM, Fann SA, Morrison JE, Lisk JR. Determining personal talents and behavioral styles of applicants to surgical training: a new look at an old problem, part I. *J Surg Educ*. 2011;68(6):534-541. DOI:10.1016/j.jsurg.2011.05.016

41. Bowe SN, Weitzel EK, Hannah WN Jr, et al. Introducing a Novel Applicant Ranking Tool to Predict Future Resident Performance: A Pilot Study. *Mil Med*. 2017;182(1):e1514-e1520. DOI:10.7205/MILMED-D-15-00436

42. Hemaida RS, Kalb E. Using the analytic hierarchy process for the selection of first-year family practice residents. *Hosp Top*. 2001;79(1):11-15. DOI:10.1080/00185860109597894

43. Forman EH, Gass SI. The analytic hierarchy process—an exposition. *Operations research*. 2001;49(4):469-486. DOI:10.1287/opre.49.4.469.11231

44. McGuire KJ, Sparks MB, Gitajn IL, et al. Selecting Residents for Predetermined Factors Identified and Thought to be Important for Work Performance and Satisfaction: A Methodology. *J Surg Educ*. 2019;76(4):949-961. DOI:10.1016/j.jsurg.2019.02.004

45. Quintero AJ, Segal LS, King TS, Black KP. The personal interview: assessing the potential for personality similarity to bias the selection of orthopaedic residents. *Acad Med*. 2009;84(10):1364-1372. DOI:10.1097/ACM.0b013e3181b6a9af

46. Zardouz S, German MA, Wu EC, Djalilian HR. Personality types of otolaryngology resident applicants as described by the Myers-Briggs Type Indicator. *Otolaryngol Head Neck Surg*. 2011;144(5):714-718. DOI:10.1177/0194599810397793

47. Boyd R, Brown T. Pilot study of Myers Briggs Type Indicator personality profiling in emergency department senior medical staff. *Emerg Med Australas*. 2005;17(3):200-203. DOI:10.1111/j.1742-6723.2005.00723.x. PMID: 15953219.

48. Hojat M, Nasca TJ, Magee M, et al. A comparison of the personality profiles of internal medicine residents, physician role models, and the general population. *Acad Med*. 1999;74(12):1327-1333. DOI:10.1097/00001888-199912000-00017

49. Schnapp BH, Ritter D, Kraut AS, Fallon S, Westergaard MC. Assessing Residency Applicants' Communication and Professionalism: Standardized Video Interview Scores Compared to Faculty Gestalt. *West J Emerg Med*. 2019;20(1):132-137. DOI:10.5811/westjem.2018.10.39709.

50. Shebrain S, Arafeh M, Munene G, et al. The role of academic achievements and psychometric measures in the ranking process. *Am J Surg*. 2019;217(3):568-571. DOI:10.1016/j.amjsurg.2018.09.015

DOI: dx.doi.org/10.5195/jmla.2025.2030

51. Stephenson-Famy A, Houmard BS, Oberoi S, Manyak A, Chiang S, Kim S. Use of the Interview in Resident Candidate Selection: A Review of the Literature. *J Grad Med Educ*. 2015;7(4):539-548. DOI:10.4300/JGME-D-14-00236.1

52. Thomas CL. 1999 op. cit.

53. Butler PD, Aarons CB, Ahn J, et al. Leading From the Front: An Approach to Increasing Racial and Ethnic Diversity in Surgical Training Programs. *Ann Surg*. 2019;269(6):1012-1015. DOI:10.1097/SLA.0000000000003197

54. Wusu MH, Tepperberg S, Weinberg JM, Saper RB. Matching Our Mission: A Strategic Plan to Create a Diverse Family Medicine Residency. *Fam Med*. 2019;51(1):31-36. DOI:10.22454/FamMed.2019.955445

55. Schenker ML, Baldwin KD, Israelite CL, Levin LS, Mehta S, Ahn J. Selecting the Best and Brightest: A Structured Approach to Orthopedic Resident Selection. J Surg Educ. 2016 Sep-Oct;73(5):879-85. DOI: 10.1016/j.jsurg.2016.04.004. Epub 2016 May 24. PMID: 27230568.

56. Martin M, Salzberg L. Resident characteristics to evaluate during recruitment and interview: a Delphi study. Educ Prim Care. 2017 Mar;28(2):81-85. DOI: 10.1080/14739879.2016.1266696.

57. Martin M, Salzberg L. Ibid.

58. Aibana O, Swails JL, Flores RJ, Love L. Bridging the Gap: Holistic Review to Increase Diversity in Graduate Medical Education. *Acad Med*. 2019;94(8):1137-1141. DOI:10.1097/ACM.0000000000002779

59. Garrick JF, Perez B, Anaebere TC, Craine P, Lyons C, Lee T. The Diversity Snowball Effect: The Quest to Increase Diversity in Emergency Medicine: A Case Study of Highland's Emergency Medicine Residency Program. *Ann Emerg Med*. 2019;73(6):639-647. DOI:10.1016/j.annemergmed.2019.01.039

60. Spottswood SE, Spalluto LB, Washington ER, et al. Design, Implementation, and Evaluation of a Diversity Program for Radiology. *J Am Coll Radiol*. 2019;16(7):983-991. DOI:10.1016/j.jacr.2018.12.00

Supplemental Files

- **Appendix A**: Methods

AUTHORS' AFFILIATIONS

Gena C. Dunivan, MD, gcdunivan@uabmc.edu, http://orcid.org/0009-0005-9021-539, Professor of OBGYN, Director, Division of Urogynecology and Reconstructive Pelvic Surgery, University of Alabama at Birmingham, Birmingham, AL; When at UNM: Gena C Dunivan, MD, Associate Dean of Graduate Medical Education, Associate Professor of OBGYN, Urogynecology and Reconstructive Pelvic Surgery Fellowship Director, University of New Mexico, Albuquerque, NM

Jonathan D. Eldredge, MLS, MA, PhD, jeldredge@salud.unm.edu, http://orcid.org/0000-0003-3132-9450, Professor, Health Sciences Library and Informatics Center and Family & Community Medicine, School of Medicine; Associate Program Director, Clinical Informatics, Graduate Medical Education, University of New Mexico, Albuquerque, NM

Marlene P. Ballejos, PhD, MPA, mballejos@salud.unm.edu, http://orcid.org/0000-0002-6288-1849, Associate Professor in Family & Community Medicine and Assistant Dean for Admissions, University of New Mexico School of Medicine, Albuquerque, New Mexico

Melissa Gonzales, MS, PhD, Mgonzales1@tulane.edu, http://orcid.org/0000-0002-1883-1967, Professor and Chair. Department of Environmental Health Sciences, Celia Scott Weatherhead School of Public Health & Tropical Medicine, Tulane University, New Orleans, LA; When at UNM: Professor, Division of Epidemiology, Biostatistics and Preventive Medicine, Associate Director Preventive Medicine Residency Program, Associate Vice Chancellor for Research and Evaluation Office of Diversity Equity & Inclusion, University of New Mexico, Albuquerque, NM

Valerie Romero-Leggott, MD, VRomero@salud.unm.edu, http://orcid.org/0000-0003-0073-8438, Vice President and Executive Diversity, Equity & Inclusion Officer; HSC Endowed Professorship for Equity in Health; Professor of Family & Community Medicine. Executive Director, UNM SOM Combined BA/MD Degree Program; UNM Health Sciences Center Office for Diversity, Equity & Inclusion.

Received October 2023; accepted April 2024

ISSN 1558-9439 (Online)

DOI: dx.doi.org/10.5195/jmla.2025.2065

ChatGPT, Python, and Microsoft Excel.

ChatGPT (version 4.0, March 14, 2024). OpenAI, San Francisco, CA, USA. https://chat.openai.com; free and subscription plans available.

Python (version 3.12.1, October 2, 2024). Python Software Foundation, Beaverton, OR, USA. https://www.python.org; free, open-source.

Microsoft Excel (version 365). Microsoft Corporation, Redmond, WA, USA. https://www.microsoft.com/excel; proprietary software, subscription-based.

The introduction of ChatGPT to the public marked a transformative shift in productivity and workplace automation, with its long-term impact yet to fully unfold. The medical library sector is no exception: in an age where effective information management and rapid access to relevant data are essential, librarians can leverage Artificial Intelligence (AI) to streamline, expedite, and enhance their daily tasks [1]. These include cataloging and organizing information, information retrieval, supporting clinical decision-making, and managing institutional knowledge, all benefiting from AI's ability to provide an additional layer of review and efficiency.

Librarians are already well-versed in using tools like Microsoft Excel to manage data and organize information. While tools like ChatGPT and Python might initially seem more complex, they are designed to be user-friendly and can greatly boost productivity when incorporated into daily workflows. Python, now integrated natively into Excel, offers a simple yet powerful coding language that enables medical librarians to automate tasks and perform advanced data analysis without leaving their familiar spreadsheet environment. The combination of Excel and Python can empower librarians to handle larger datasets, automate repetitive tasks, and generate more sophisticated insights with ease.

The integration of AI into professional routines is reshaping expectations across industries. As AI tools become more widespread, there is a growing assumption that leveraging these technologies for data analysis, decision-making, and automation is now a standard part of "doing good work." In the medical library field, this means that harnessing AI and Python for international collaboration, managing vast amounts of data, and providing rapid, accurate insights is becoming not just an enhancement, but a crucial component for maintaining excellence in research and information management.

CHATGPT

ChatGPT, a variant of the Generative Pretrained Transformer (GPT) models developed by OpenAI, exemplifies the rapid advancement in natural language processing (NLP) [2]. It leverages deep learning algorithms to understand context and generate coherent, contextually appropriate responses, marking significant milestones in AI development [3]. The first version of GPT was launched in 2018, and the model has since evolved into increasingly sophisticated conversational AI. The latest version, GPT-4, introduced in 2023, enhances these capabilities further, with multimodal abilities to process both text and images. As a result, ChatGPT has quickly become a transformative force across diverse industries, showcasing its adaptability and robust capability in enhancing operational performance and customer interactions [4].

PYTHON

Python is a high-level programming language renowned for its clear syntax and readability, making it an ideal choice for both novice and experienced programmers. Developed by Guido van Rossum and first released in 1991, Python has evolved significantly over the years, becoming one of the most popular programming languages globally. Its versatility, coupled with extensive community support, has contributed to its widespread adoption across industries. As technology continues to advance, Python's role in developing applications that process and analyze complicated sets of data is expected to expand, securing its position as a leading tool for technological innovation in the years ahead.

MICROSOFT EXCEL

Since its launch in 1985, Microsoft Excel has evolved from a basic spreadsheet tool into a comprehensive platform for data analysis, visualization, and automation. Originally intended for simple calculations, Excel has progressively incorporated advanced features such as pivot tables, intricate formulas, and Visual Basic for Applications (VBA), empowering users to automate tasks and develop custom functions. More recently, the integration of Python formulas in Excel represents a significant leap forward, enabling users to harness powerful data science tools and perform advanced analyses directly within Excel. This evolution underscores Excel's adaptability and its continued relevance in the dynamic, data-driven decision-making environment.

RESEARCH OBJECTIVE

This article illustrates how ChatGPT can serve as a valuable digital assistant for medical librarians, offering:

(1) support in providing second opinions for their daily decisions,

(2) the ability to generate user-friendly Python scripts that automate repetitive tasks, particularly when processing and analyzing data collection such as Excel spreadsheets,

DOI: dx.doi.org/10.5195/jmla.2025.2065

(3) tools to quickly summarize and synthesize relevant information from large volumes of data or literature, and

(4) the capability to translate or summarize information from languages in which the librarian may not be proficient.

The primary goal is to offer practical solutions that can significantly enhance the competence and effectiveness of medical librarians in their day-to-day responsibilities.

INTEGRATED TOOL APPROACH

ChatGPT has emerged as a powerful tool for enhancing the productivity of information and data management by leveraging its capabilities with Excel spreadsheets and Python coding. By automating repetitive tasks, generating custom scripts, and providing real-time assistance in data processing, ChatGPT accelerates the construction, analysis, and synthesis of complex sets of information [5]. Its ability to seamlessly integrate with Excel allows users to manipulate and visualize data with ease, while its proficiency in Python enables the execution of advanced algorithms and data analysis techniques.

This combination not only streamlines workflows but also empowers users to derive insights more rapidly, making ChatGPT an invaluable asset in data-driven environments. For example, to have data entry automation, the combination of ChatGPT with Excel spreadsheets automates repetitive data entry tasks, and with Python programming generates scripts to automate data entry processes. This could be extended to a lot of other potential needs, such as data analysis, data visualization, formula assistance and data cleaning, integrating these possibilities as an automated analytical solution to manage and understand a database, as exemplified in Figure 1.

ChatGPT's ability to integrate with tools like Excel and Python has become indispensable in today's data-driven landscape. As the volume of information continues to grow, this capability empowers users to effectively organize and interpret sophisticated datasets while seamlessly integrating them [6]. By leveraging Excel for data organization and visualization, combined with Python's powerful coding features for advanced analysis and automation, ChatGPT facilitates faster and more accurate decision-making, enhancing productivity across various tasks.

ANALYZING FROM A MEDICAL LIBRARIAN'S PERSPECTIVE

The adoption of ChatGPT and its ability to work with Excel and Python in the operational frameworks of medical libraries represents a pivotal enhancement in the methodologies for managing and analyzing information. ChatGPT is highly effective in assisting with data organization, analysis, and visualization in Excel, making it easier for librarians to handle elaborate data compilation and respond to user queries. Python, generated by ChatGPT, provides robust support for automating repetitive tasks and performing advanced data analysis, optimizing the storage, retrieval, and synthesis of information – that can be copy and pasted in Excel easily (Figure 1). Together, these tools create a comprehensive system that significantly improves the efficiency and effectiveness of library services, empowering medical librarians to manage the growing complexity of data in their daily work.

The integration of ChatGPT with Excel and Python offers a powerful boost to medical library operations (Table 2). With ChatGPT's natural language processing, medical librarians can easily manage queries and guide users through difficult data tasks. By utilizing Excel for data organization and Python for automation, they can streamline processes, improve accuracy, and enhance the support provided to healthcare professionals and researchers.

Figure 1 Using ChatGPT to create Python scripts

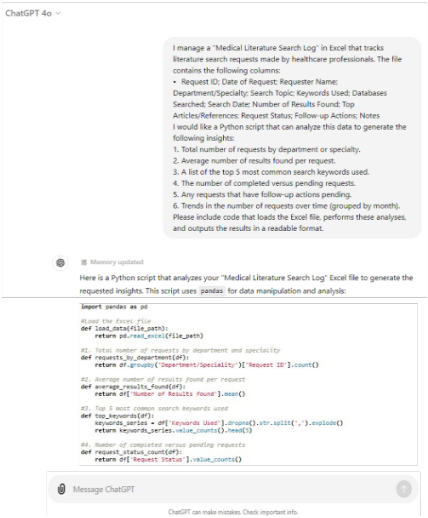


Table 2 Common use cases for using ChatGPT, Excel, and Python combined

Use Case	Role of ChatGPT	Role of Excel	Role of Python	Combined Impact
Information Retrieval	Helps formulate precise search queries and strategies	Organizes and retrieves data from extensive information set	Automates the extraction of data from various sources	Faster and more accurate retrieval of relevant information
User Interaction	Engages with users to clarify needs and provide tailored responses	Displays and organizes user-requested data in a clear format	Automates user queries and data processing tasks	Enhanced user experience through efficient, responsive interactions
Data Management	Assists in structuring and categorizing large sets of data	Manages data organization using tables, filters, and pivot tables	Automates data management processes, including cleaning and integration	Streamlined data management with reduced manual effort and increased accuracy.
Research Assistance	Provides guidance on research methodologies and data analysis	Facilitates data analysis and presentation through charts and reports	Executes advanced data analyses and simulations	Comprehensive support for research, leading to deeper insights and quicker results
Educational Support	Offers instructional content and explanations for data-related tasks	Organizes educational materials and examples in a structured format	Automates the generation of educational data repository and simulations	Enhanced learning experience with interactive, data-driven resources

DOI: dx.doi.org/10.5195/jmla.2025.2065

However, adopting ChatGPT, Excel, and Python also comes with challenges. These include a learning curve, adapting workflows, and ensuring smooth integration with existing systems. Data security and privacy are additional concerns that require careful attention. Addressing these issues through proper training, infrastructure investment, and proactive management is key to unlocking the full potential of these tools and improving library services.

CONCLUSION

This article demonstrates how the combination of ChatGPT, Excel, and Python can greatly enhance the productivity of medical librarians by automating repetitive tasks and improving data management processes. By utilizing ChatGPT's ability to interact with Excel spreadsheets and generate Python code, medical librarians can competently analyze and handle large datasets, streamlining their daily operations. The exploration of these technologies underscores their transformative potential in boosting effectiveness, accuracy, and the personalization of library services. However, successful integration requires careful planning and management. Looking ahead, ongoing advancements in AI and programming promise even greater improvements in productivity, enabling librarians to take on more strategic roles supported by technology.

REFERENCES

1. A. Spencer and J. Eldredge, "Roles for librarians in systematic reviews: A scoping review," Journal of the Medical Library Association, vol. 106, no. 1, 2018. DOI: 10.5195/jmla.2018.82.
2. P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," Internet of Things and Cyber-Physical Systems, vol. 3, pp. 121–154, 2023. DOI: 10.1016/j.iotcps.2023.04.003.
3. M. Sallam, "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns," Healthcare, vol. 11, no. 6, p. 887, 2023. DOI: 10.3390/healthcare11060887.
4. A. Loxley, ChatGPT Millionaire Bible: Unleash Your Potential and Boost Your Success in the Digital Age Mastering the Power of ChatGPT, Independently published, 2023.
5. E. A. M. Van Dis, J. Bollen, W. Zuidema, R. Van Rooij, and C. L. Bockting, "ChatGPT: five priorities for research," *Nature*, vol. 614, no. 7947, pp. 224–226, Feb. 2023, DOI: 10.1038/d41586-023-00288-7. PMID: 36737653.
6. H. Kissinger, E. Schmidt, and D. Huttenlocher, The Age of AI: And Our Human Future, Back Bay Books, New York, 2022.

Kaique Sbampato,
kaique.sbampato@poli.ufrj.br,
https://orcid.org/0000-0002-3287-9318,
Production Engineering Program, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.

Humberto Arruda,
Humberto.arruda@poli.ufrj.br,
https://orcid.org/0000-0002-3064-809X,
Production Engineering Program, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.

Édison Renato Silva, *edison@poli.ufrj.br,*
https://orcid.org/0000-0002-4994-7057,
Industrial Engineering Department and Production Engineering Program, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.

ISSN 1558-9439 (Online)