REPORT

2024

Implementation of Machine Learning in Cluster for Reviews and Health Technology Assessments: Results for ML 3.0

NIPH
Norwegian Institute of Public Health

Institution	Norwegian Institute of Public Health Guri Rørtveit, Director-General
English title	Implementation of Machine Learning in Cluster for Reviews and Health Technology Assessments: Results for ML 3.0
Authors	Tiril C. Borge, Researcher Heather Melanie Ames, Senior Researcher Patricia Jacobsen Jardim, Senior Adviser Hans Bugge Bergsund, Researcher Martin Smådal Larsen, Adviser Christopher Rose, Researcher
ISBN	978-82-8406-441-3
Publicatin type	Report
Number of pages	24
Commissioned by	Norwegian Institute of Public Health
Key words (MeSH)	biomedical; artificial intelligence, health; unsupervised machine learning; supervised machine learning; deep learning
Citation	Borge TC, Ames HM, Jardim PJ, Bergsund HB, Larsen MS, Rose C, (2024). "Implementation of Machine Learning in Cluster for Reviews and Health Technology Assessments: Results for ML 3.0" [Implementering av maskinlæring i klynge for vurdering av tiltak: Resultater for ML 3.0]. Oslo: Folkehelseinstituttet 2024.

Contents

CONTENTS 2

KEY MESSAGES 3

PREFACE 4

BACKGROUND 5

ML TEAM 3.0 DELIVERABLES 7

FINAL REFLECTIONS 21

REFERENCES 23

Key Messages

In 2020, the Cluster for Reviews and Health Technology Assessments (HTV) at the Norwegian Institute of Public Health (NIPH) established a dedicated machine learning (ML) team. The ML team has since become an international leader in integrating and implementing ML into evidence synthesis, achieving significant milestones, and securing official financing in November 2022, which contributed to much of the ML activities performed by ML 3.0.

The overall goal of the ML team is to use ML in a way that best combines human intelligence and ML, to enhance human activities, by figuring out how best to integrate ML and workflow changes, throughout the review process.

This report outlines the team's activities during its iteration, ML 3.0, covering implementation, peer-to-peer support, dissemination, evaluations, innovation, horizon scanning, and external networking and collaborations.

ML Team 3.0 accomplished a variety of project deliverables, including providing ML support to six teams, conducting teaching sessions, implementing an ML reporting template, and implementing e-learning course. Dissemination efforts included presentations, poster sessions, and publications, while evaluations encompassed various projects, including a pilot on interrater agreement using ChatGPT. Innovations comprised development of a scalable e-learning course, a survey on ML attitudes and barriers, and qualitative interviews.

Title:
Implementation of Machine Learning in Division for Health Services: Results for ML 3.0

Publisher:
The Cluster of Reviews and Health Technology Assessments, Division for Health Services, ML team 3.0

Type of publication:
Report

Preface

This report presents results for the current iteration of the ML team, “ML 3.0”.

Financing
Much of the work, particularly relating to innovation activities, was externally funded. The remaining work was self-initiated and financed by the Cluster for Reviews and Health Technology Assessments, Division for Health Services at the NIPH.

With appreciation
The current team’s learning and strategizing are due not only to the dedication of its members, past and present, but also to HTV leadership’s investment and vocal support. There have also been numerous colleagues who have provided support, feedback, ideas, and opportunities including the team of librarians who have started evaluation work on OpenAlex. Outside of NIPH, James Thomas’ mentoring and his team at EPPI Centre have continued to be instrumental to our understanding of ML and its potential to provide the most valuable evidence synthesis products to our commissioners.

Conflicts of interest
All authors declare they have no conflicts of interest.

Kåre Birger Hagen	Rigmor C Berg	Tiril Borge	Heather Ames
Research director	*Department director*	*Project leader*	*Implementation leader*

Background

Since early 2020, the Cluster for Reviews and Health Technology Assessments (HTV) at the Norwegian Institute of Public Health (NIPH) recognized the potential benefits of employing machine learning (ML) in evidence syntheses. Consequently, a dedicated ML team was funded in late 2020, aligning with NIPH strategies for 2019-2024 focused on automation and workflow innovation.

Since its inception in late 2020, the ML team has positioned NIPH as a leader in integrating and implementing ML into evidence synthesis, strategically innovating to ensure a sustainable competitive advantage.

ML optimally utilizes scarce human resources by handling complex, repetitive tasks, allowing reviewers to focus on more thought intensive parts of the evidence synthesis process. The ML Team 2.0 secured official funding in November 2022, which has funded much of the ML related work in HTV during 2023, particularly in relation to implementation.

In this report we will present ML 3.0s deliverables, in relation implementation, dissemination, evaluations, capacity building and external networking and collaborations.

Goals

The overall goal of the ML team is to use ML in a way that best combines human intelligence and ML, to enhance human activities, by figuring out how best to integrate ML and workflow changes, throughout the review process. In 2023, due to workforce reduction at the beginning of the year, and external financing for implementation activities, ML team 3.0 has focused on capacity building activities as well as any already started evaluation activities, while horizon scanning, and innovation activities have largely been put on hold.

Specific goals for ML 3.0, based on the team announcement in 2022 and goals put forward as part of the externally funded activities, have been:

1. To scale up an agile, sustainable and evidence-based ML skills development programme, involving 4 sub-goals:
    a. Capacity building of conceptual aspects of three ML functions through e-learning courses

   b. Explore attitudes, barriers, experiences around the use of ML among employees
   c. Investigate approaches to changes in work processes related to the use of ML.
   d. Develop an implementation guide for internal and external use.
2. To continue to facilitate all review teams having the knowledge and confidence to use ML functions in their evidence synthesis work
3. Disseminate the results and experiences of the ML team's work in the form of an implementation guidance document
4. Continue with already commenced evaluation activitites
5. Maintain external networking and collaboration activities

ML team 3.0 deliverables

The following text details ML 3.0 team activities undertaken from January 2023 up until end of March 2024, which is when the external funding period ended. This section reports on all activities and their quantitative and qualitative results. We present activities specifically related to each goal, with the first four subheadings relating to the externally funded activities.

1a. Capacity building of conceptual aspects of three ML functions through e-learning courses

Develop scalable e-learning.

Background: In March 2022, we piloted a one-week intensive training programme. The pilot included digital workshops that built theoretical and practical knowledge. Despite the value of the training programme, several improvements were seen as needed:

a) increase transferability to our colleagues in other divisions and departments, where knowledge products are produced for different users and possibly with different processes.
b) compress the content so that it is feasible for new and existing employees in less time.
c) transform the way training is delivered so that it is scalable both in parts and as a whole, from live teaching to recorded videos, interactive training and tutorials, quizzes and independent tasks/exercises.

Creating an effective and engaging e-learning course involved careful consideration of various design aspects. Clear learning objectives were defined to ensure that the course content and activities aligned with employees' specific needs. Personalization features allow learners to highlight their skills and bypass content they already know, catering to novices by starting with foundational concepts and gradually progressing to advanced topics as confidence grows, while intermediate and advanced learners can concentrate on more advanced modules. Emphasis was placed on using plain and easy-to-understand language throughout the course. Assessments and feedback mechanisms are integrated into the course to help learners monitor their progress and pinpoint areas for improvement.

We developed e-learning modules on our most used ML functions: Ranking algorithms, classification algorithms and OpenAlex. The e-learning content goes through the conceptual aspects of these functions, to provide an understanding of what they are and how they work, as this is fundamental to understanding when it is appropriate to use them and how to use them correctly.

ML week 2023

In November three representatives from the EPPI centre held in-house training sessions on the EPPI Reviewer tool and our most used ML functions. Forty-nine employees participated in an introductory workshop, where most participants (39 out of 49) were from other departments in NIPH including: Norwegian Scientific Committee for Food and Environment (VKM), Global Health, Centre for Epidemic Interventions Research (CEIR), Chemical Toxicology, Air Quality and Noise, Norwegian Poison Information Centre, Library at NIPH (Biblioteket) , Antibiotic Resistance and Infection Prevention, Helsebiblioteket.no, Mental Health and Suicide, Food Safety, Childhood and Families and Physical Health and Ageing. A prerequisite to participate in the intermediate sessions on ML functions during ML week was to finish the e-learning course. The advanced workshops were attended mostly by HTV employees.

1b. Explore attitudes, barriers, experiences around the use of ML among employees

Development of survey

In May, four members of the team started developing a survey to assess attitudes, barriers and experiences concerning the use of ML. This would serve the dual purpose of both evaluating the results of capacity-building so far in HTV, as well as informing the development of the implementation strategy. The process of developing the survey was conducted in four stages. First, an OpenAlex-based search of qualitative literature on ML experiences was conducted to inform the content of the questionnaire. Second, another OpenAlex-based search of studies describing relevant AI/ML attitude questionnaires was conducted, with the purpose of developing a pool of AI/ML-attitude questionnaire items. Third, items that corresponded to the most salient themes in the qualitative papers were drawn from the pool and used to develop our own AI/ML-attitude survey. Finally, a selection of our peers (members of the team, external evidence synthesis professionals and union representatives) were asked to review the survey and provide feedback (face validation). In response to the feedback, we created some additional items in areas that were not covered by the existing item pool and adjusted the way in which some of the items were phrased.

The survey has been distributed twice so far in HTV, once before and once after the capacity building activities conducted outside the team in November. Analysis is under way and will be completed during Q1 2024.

Qualitative interviews by Comte Bureau

Comte Bureau was commissioned by NIPH to conduct ten qualitative interviews with NIPH employees and leaders to understand the attitudes, barriers, and experiences around the use of ML in HTV. Prior to the interviews, three meetings were held in the autumn of 2023 to review the preparations for the interviews and lessons learnt from the parallel quantitative survey. The research questions from which the interview guide was developed were as follows:

1. What attitudes, barriers and experiences do FHl employees have to ML?
2. What personal barriers occur?

3. What structural barriers occur in relation to using ML in current projects in NIPH?
4. What does it take for NIPH employees to use ML in their everyday work?
5. What are NIPH employees' experiences of using ML with regard to
    a. How they have used ML so far
    b. How have they experienced work process changes associated with ML (if they have?)
    c. Perceived change in job satisfaction or engagement
6. How have employees experienced the capacity building activities that have occurred since September 2023 (e-learning course, ML week, change workshop by Mindshift at HTV seminar in October)?
7. How have managers facilitated the use of ML?

The results from the interviews are published in a report.

1c. Investigate approaches to changes in work processes related to the use of ML

Scrum master course

In July, three members of the team participated in a Scrum master certification course. The aim with Scrum was originally to establish a more effective approach to software development, challenging traditional methods deemed dysfunctional, however the Scrum methodology can be applied across many different subject areas. Scrum is grounded in empirical process control theory, emphasizing a clear vision and learning through short iterations toward the goal. Work is conducted by small, self-organized teams, focusing on solving problems as they arise. The Scrum theory methodology emphasize transparency, inspection, and adaptation. The framework is minimalist, allowing Scrum teams to fill in details of their chosen methodology, fostering strong ownership and results. Overall, the Scrum methodology's emphasis on adaptability, collaboration, and continuous improvement makes it a valuable approach for teams working in dynamic and complex environments, where changes are frequent, and requirements evolve over time.

Digital learning institute – Digital design diploma

As the ML team's knowledge on e-learning development was very limited, three of the team members registered for A Professional Diploma in Digital Learning Design. The course content included learning about the foundations of instructional design, encompassing an understanding of learning theories, instructional design models, and the ability to analyse learner needs to formulate learning objectives. The program also explores digital learning technologies, including various tools and technologies utilized in online learning, and provides insights into Learning Management Systems (LMS) and content authoring tools. Content development is a key focus, involving the design and creation of engaging and interactive digital content, incorporating multimedia elements such as videos and interactive modules. There is focus on user-friendliness and accessibility, incorporating principles of design thinking into course development. Additionally, the program addresses adaptive learning and personalization techniques, including how to create personalized learning experiences and implement adaptive learning strategies based on learner needs. Effective communication strategies for online learning and fostering collaboration among learners in a digital environment are also integral components of the diploma program.

Change workshop by Mindshift at cluster seminar.

The use of ML and AI in our work processes inherently causes a change in how we work. Additionally, HTV has been through major downsizing effectuated in early 2023 and a major reorganization effectuated in January 2024, and we therefore saw a need for increasing the employee's knowledge in handling change. Therefore, during HTV seminar in October, we commissioned Mindshift to hold a workshop on change processes with a focus on digitalisation, where the main aim was to increase employees understanding of how to manage change as part of everyday working life. Part of the workshop was held separately for leaders and the remaining employees. The leaders' course covered topics such as the role of leaders in change processes, communication with employees, and guidance on preserving employee well-being. The employee course followed a similar structure, covering topics like self-management, handling changes in the work environment, and stress and stress management. Both courses were facilitated via practical tasks, plenary discussions, and feedback sessions.

1d. Develop an implementation guide for internal and external use

Background and aim

As part of external financing secured in late 2022 an implementation guidance is to be developed, based on the ML teams experiences with implementation since the team's inception in 2020. The implementation guide will be tailored for evidence synthesis groups or institutions aiming to introduce ML into the evidence synthesis process within a ML-naïve work environment. Even though most of our implementation work has been centred around ML functions related to literature searching and screening, the guide aims to be sufficiently general for implementation of various ML functions and tools beyond the presented examples. Drawing from our implementation experiences, the document offers a practical framework for ML implementation into evidence synthesis processes within research institutions, serving as a roadmap adaptable to each institution's organizational goals and objectives.

Content

The implementation guide will be divided into three main sections reflecting the phases of implementation: pre-implementation, implementation, and sustainment/ evaluation. In each section we will present key tasks that need to be conducted, discuss the roles different actors will play and suggest ways in which support can be given and feedback gathered. Throughout the report we will provide general guidance based on what we have learned during our implementation process.

2. To continue to facilitate review teams having the knowledge and confidence to use ML functions in their evidence synthesis work

One prioritized task of ML 3.0 outside the activities related to the external funding, was to support review teams in implementing ML in their evidence synthesis work, which mainly involved peer-to-peer support to individual teams. We supported both teams in HTV as well as teams in other areas of NIPH that were conducting evidence syntheses. Table 1 details team

activities related to peer-to-peer support undertaken. We have provided ML support to six teams across the institute, as well as holding ML teaching sessions at STAMI. We have also implemented the updated reporting template for ML use in our reports and an e-learning course to foster conceptual understanding of our most used ML functions.

Table 1: *Peer-to-peer support activities*

Date	Title	Type of deliverable
January 2023	Implementing the new ML template for our systematic reviews and scoping reviews. These were developed, peer reviewed, and pilot tested before they were implemented	Implementation template
January 2023	Help to report ML in « Triclosan coated sutures for prevention of surgical site infection: a health technology assessment»	Peer-to-peer support
March-May 2023	ML education for STAMI, and support to use EPPI. Also contributed to implementing ML in their systematic review protocol	ML teaching sessions at STAMI
March-May 2023	ML support to "A knowledge resource for municipalities"	Peer-to-peer support
April-May 2023	General EPPI Support to Norwegian Scientific Committee for Food and Environment (VKM), answering EPPI related questions and supporting the set-up of EPPI-VIS.	Peer-to-peer support
May-September 2023	ML support for global health for a qualitative evidence synthesis entitled "Using evidence from civil society in health policy processes: a qualitative evidence synthesis"	Peer-to-peer support
October 2023	ML support for the project "CFS/ME IPD screening"	Peer-to-peer support
November 2023	Implementing a conceptual e-learning with four modules for the most used ML algorithms/functions in EPPI Reviewer	E-learning
November 2023 - February 2024	ML support to "New national guidance on new recommendations for screening of resistant microbes", Division of Infection Control	Peer-to-peer support

In table 2 we provide a list of reports either published or completed in 2023 that have reported the use of at least one ML function. There might be other publications, like "forskningsomtaler", that have used ML but due to the individual publications word limitations, have not described their use of ML. ML support was provided to all but one of the reports presented in table 2.

Table 2: *Internal reports using ML, published, or completed*

English Title	ML functions used	Provided ML support?
Children and young people who perpetrate serious acts towards others: a rapid review	OpenAlex, Priority screening	Yes*
Children and young people's opinions on topics in the proposal for a new Children's Act: a scoping review of Nordic qualitative studies	Priority screening, Cochrane RCT classifier, Clustering	Yes*
The use of force and limit-setting for children and youth in residential childcare and foster care: systematic scoping review (update)	OpenAlex, Priority screening	Yes*
Portable ECG equipment for the diagnosis of atrial fibrillation in the specialist health care: rapid HTA- scoping review	Priority screening	No
Co-therapy and reflecting teams in couples- and family therapy: a mixed methods systematic review	Priority screening	Yes*
Parental follow-up in family welfare services after child removal: a scoping review	Priority screening, clustering	Yes*
What contributes to stable placements when children are placed in foster homes or institutions? Systematic literature search with sorting	OpenAlex, Priority screening	Yes*
Individual placement and support for people with moderate to severe mental illnesses or substance abuse disorder: a systematic review	Priority screening	Yes*
What are the characteristics of youth who are placed in care institutions in child welfare? A rapid review	OpenAlex, Priority screening	Yes*
Consequences of the Covid-19 pandemic on children and youth's life and mental health: Second update of a rapid review	OpenAlex, Priority screening, custom classifier	Yes*
Post COVID-19 condition after Omicron: a rapid review	Priority screening	Yes*
Surgery for degenerative rotator cuff tears: a health technology assessment	Priority screening, custom classifier	Yes
Transcutaneous non-invasive vagus nerve stimulation (gammaCore) for the treatment of cluster headache: A single technology assessment	Open Alex, Priority screening, Cochrane RCT classifier	Yes
Triclosan coated sutures for prevention of surgical site infection: a health technology assessment	Cochrane RCT classifier, Economic evaluation	Yes

	classifier, Priority screening	
Coercion in mental health care and violence: systematic literature search with sorting	Priority screening	Yes*
Language screening tools for children 0-5 years: a systematic scoping review	Priority screening, OpenAlex, custom classifier, clustering	Yes*

**A member of the ML team was also a member of the project team*

3. Disseminate the results and experiences of the ML team's work

The team has been highly active in 2023 disseminating our work, mainly outside NIPH. We have held one presentation in-house and fourteen presentations outside of NIPH, where four were outside of Norway. We have had four poster presentations. A summary of all our dissemination activities is found in Table 3.

Presentations and posters

Table 3: *Completed dissemination activities*

Date	English Title/Description	Type
March 2023	Implementation and evaluation activities to build support for machine learning in evidence syntheses	Poster, NIPH Research and innovation day
May 2023	Managing the information explosion: the usefulness of machine learning in SIA – the NIPH example	Presentation, INSIA annual meeting, Stockholm
August 2023	Machine learning versus automation in evidence syntheses	Presentation, NORNESK webinar
September 2023	Use of artificial intelligence and machine learning in evidence syntheses	Presentation, Division seminar
September 2023	Can using the Cochrane RCT classifier help speed up study selection in qualitative evidence syntheses (QES)? A retrospective evaluation	Poster Cochrane Colloquium, London
September 2023	Building acceptance for machine learning in study selection within a systematic review institution: Experiences from the Norwegian Institute of Public Health	Long Oral Presentation, Cochrane Colloquium, London
September 2023	Connecting with other researchers who are working with ML or who want to work with ML. Networking for future collaborations.	Cochrane networking
September 2023	NIPHs most frequently used machine learning functions in evidence syntheses	Presentation, NORNESK webinar

September 2023	Information about ML team and our work published on the NIPH webpage about AI at FHI	NIPH AI webpage
September 2023	Implementation and evaluation activities to build support for machine learning in evidence syntheses	Poster, Public Health Conference in Tromsø
September-October 2023	To make the ML resources available on our SharePoint site more appealing and inviting, and more intuitive with regards to where you can find the different resources available.	Create new SharePoint site
October 2023	How can machine learning be used to keep you updated on research areas?	Presentation, NORNESK webinar
October 2023	Use of artificial intelligence and machine learning in evidence syntheses	Presentation, VKM
October 2023	Building acceptance for machine learning in study selection within a systematic review institution: Experiences from the Norwegian Institute of Public Health	Presentation, Will Moy (Campbell Collaboration) & UK cabinet office
November 2023	How to implement machine learning in evidence syntheses	Presentation, NORNESK conference, Bergen
November 2023	How does the use of machine learning in evidence syntheses affect our work processes?	Presentation at NORNESK conference, Bergen
November 2023	Experiences with the use of artificial intelligence and machine learning in evidence syntheses	Presentation, OsloMet seminar "Artificial intelligence in evidence syntheses - is OsloMet keeping up?"
January 2024	How NIPH utilize AI tools in the evidence synthesis process	Presentation, University of Oslo
January 2024	Implementation guidance	Poster, Norwegian Network for Implementation Research (NIMP) Conference
February 2024	Use of machine learning in searching screening and categorising literature	Presentation, Norwegian Poisons Information Centre
February 2024	What do we know about the effect of ML/AI for evidence synthesis in medicine and allied fields? Some challenges and opportunities for the future	Presentation, National Academy of Sciences, Engineering and Medicine, Texas A&M Institute for Advancing Health Through Agriculture
February 2024	Building acceptance for machine learning in study selection within a systematic review	Presentation, UK Health Security Agency

	institution: Experiences from the Norwegian Institute of Public Health	
February 2024	Use of machine learning in searching screening and presentation of literature	Presentation, Regional kompetansetjeneste for rehabilitering Helse Sør-Øst/ Sunnaas Rehabilitation Hospital
March 2024	Exploring the advancements of machine learning and artificial intelligence in evidence synthesis: some applications, possibilities, and challenges	Presentation, INSIA webinar
March 2024	NIPHs use of machine learning for faster project delivery	Presentation, Samarbeidskonferanse i Oslo og Viken

Publications and preprints

In 2023 we have published two protocols and one book chapter, and one paper is in review per December 2023. Additionally, two papers and one report are in progress, aimed at publication during Q1 2024. Details on the publications are listed in Table 4.

Table 4: *List of publications and preprints published or in progress*

Date	Title	Type
January 2023	**Muller, A. E., Berg, R. C., Meneses-Echavez, J. F., Ames, H. M. R., Borge, T. C., Jardim, P. S. J**., Cooper, C., & **Rose, C. J**. (2023). The effect of machine learning tools for evidence synthesis on resource use and time-to-completion: protocol for a retrospective pilot study. Systematic reviews, 12(1), 7. https://doi.org/10.1186/s13643-023-02171-y	Protocol
August 2023	**Rose, C**., Ringsten, M., Bidonde, J., Glanville, J., **Berg, R**., Cooper, C., Muller, A., **Bergsund, H**., **Meneses Echávez, J**., & Potrebny, T. (2023). Using a large language model (ChatGPT) to assess risk of bias in randomized controlled trials of medical interventions: protocol for a pilot study of interrater agreement with human reviewers. https://doi.org/10.21203/rs.3.rs-3288515/v1	Protocol
December 2023	**Ames, H. N**., Noyes, J., and A. Booth (2023). Chapter 6: Selecting studies and sampling. Draft version. Cochrane-Campbell Handbook for Qualitative Evidence Synthesis, Version 1.0. Jane Noyes (Senior Editor) and Angela Harden (Senior Editor). London, Cochrane. https://training.cochrane.org/cochrane-campbell-handbook-qualitative-evidence-synthesis	Book chapter
In review	**Meneses-Echavez, J. F., Muller, A. E., Berg, R. C., Ames, H. M. R., Borge, T. C., Jardim, P. S. J**., Cooper, C., & **Rose, C. J**. The effect of machine learning tools for evidence synthesis on resource use and time-to-completion: a retrospective pilot study.	Paper

In progress	**Bergsund, H.B., Larun, L., Lidal, I., Poulsson, A., Borge, T., Jardim, P., Ames, H**. Tentative title: Developing a questionnaire to explore attitudes towards implementation of machine learning in a systematic review setting: a worked example.	Paper
In progress	Implementation guidance – report	Report
In progress	Implementation guidance – paper	Paper
In progress	Can using the Cochrane RCT classifier help speed up study selection in qualitative evidence syntheses? A retrospective evaluation	Paper

4. Continue with already planned evaluation activities

Complete

1. We evaluated whether the Cochrane RCT classifier could be used to identify qualitative studies. It was tested on 2828 included primary qualitative studies from a total of 102 QES's. The findings were presented as a poster at Cochrane Coloquium september 2023, and will also be published as a paper.
2. We have completed an evaluation comparing reviews that used versus did not use ML with respect to resource use and time-to-completion. The manuscript was submitted to Systematic Reviews on 6/10/2023 and is awaiting peer review. The protocol for the study is published (1).
3. In May we sent out a questionnaire to all employees in HTV. The questionnaire was ment to assess the employees view on the ML functions they have used and the assistance they had received from the ML team and the ML resources. Some main findings summarized, based on open questions:
    a. **Challenges with the use of ML functions:**
    Several ed report limited knowledge and experience with ML functions, hindering independent setup and use. Some experienced the complexity and limited application areas for certain ML functions. The time intervals between use create a need for repeated training. Changes over time in understanding and use of ML functions, as well as challenges related to EPPI Reviewer's usability, are also mentioned. Discussion of methodology and risk of bias is highlighted, including the need for harmonisation with risk of bias tools.
    b. **Support needs and suggestions for improvements:**
    Employees value personalised support and guidance from ML team members. Availability and helpfulness of the ML team is positively rated. Materials and training resources work well, but some call for better identification of relevant materials. Discussions and conversations about ML functions should be increased. Resource persons with ML expertise, training and support for EPPI-Reviewer are emphasised as useful. A desire for more detailed guidance and early involvement of the ML team in projects is also mentioned.
    c. **Improvement suggestions for easier use of ML functions:**

Employees want more focus on experience sharing from others who have used ML functions. Increased user-friendliness and intuitiveness in the programmes and tools associated with ML are desired. Several suggest more training and information about new opportunities and functions, as well as clearer information about support contacts. The idea of super users and better understanding of areas of use for the municipal team (kommunelag) is emphasised, along with the need for regular use and practice with the tools to improve understanding and mastery.

Based on this feedback we identified areas where the team should put in efforts, which guided much of the work within HTV the remaining year: Enhance and maintain knowledge, increase visibility and closer follow-up. To enhance and maintain knowledge, we have developed a new e-learning course. Additionally, a new ML learning week was arranged in November, spanning three full days, aiming to increase knowledge and familiarity with EPPI Reviewer and our most used ML functions. The e-learning modules were required pre-work before the ML learning week. Some sessions in the ML week were also open to all NIPH employees. To foster closer collaboration with EPPI, they held most of the sessions during the ML week.

To increase visibility and accessibility, we have fully renovated our ML and AI SharePoint site, with the aim of being a more intuitive and user-friendly "knowledge base". The SharePoint room houses relevant ML information, including reporting guidelines, help request forms, relevant literature as well as links to the e-learning course. This centralized resource aims to empower employees to find solutions independently, fostering a self-sufficient and informed workforce. To allow for closer follow-up, we have implemented more regular feedback rounds, both officially through surveys and unofficially through conversations with employees.

Ongoing
A pilot project to evaluate interrater agreement between human consensus risk of bias judgements and those made by a large language model (ChatGPT). A protocol has been written and published as a preprint and submitted to BMC Medical Research Methodology for peer review. An international team has been formed, with researchers from Norway (NIPH, Western Norway University of Applied Sciences (HVL)), Sweden (University of Lund), and the United Kingdom (University of Bristol, York). The first phase of the work is underway, and a preprint of the protocol is published (2).

5. Maintain external networking and collaboration activities

Below we list some of the groups and institutions with whom we presently engage in various capacities, ranging from networking to collaboration. Certain affiliations are characterized by networking connections and the exchange of expertise, while others involve more substantive collaborations, including contributing data for the advancement of ML functions, providing user input on tools and ML functions, and jointly undertaking initiatives such as strategy development and the planning of forthcoming evaluations.

Campbell Collaboration

Will Moy, CEO of the Campbell Collaboration, reached out to the ML team based on recommendations from James Thomas, as they, together with representatives from the UK Cabinet Office who work on evidence and evaluation, are exploring what it would take to support a step up in the use of ML in evidence synthesis. Based on this they wanted to have our expertise and experience on board. The Campbell Collaboration, together with EPPI centre and Future Evidence Foundation are putting together a proposal related to accelerating work on automation for Evidence Synthesis, based on the need for evidence synthesis needing to be faster, cheaper, and more widely available than it is now. This will include running evaluations that the group judges as contributing useful information regarding the performance of automation tools in evidence synthesis. The ML team has expressed interest in participating, but this project is still in its infancy, and the scope of the work which the ML team will contribute to is still unclear.

Cochrane Qualitative and Implementation Methods Group (QIMG)

HA sits as a co-convenor of the QIMG. As part of her role as co-convenor she has authored a chapter in the upcoming Cochrane-Campbell Handbook for Qualitative Evidence Synthesis on study selection and sampling where the use of ML is discussed. She is also an academic editor on the handbook. Together with Prof. James Thomas, she has responsibility for raising ML concepts with the QIMG when they are relevant or answering questions about potential use of ML in qualitative evidence synthesis.

EPPI Centre and National Institute for Health Care Excellence (NICE)

The study begun in late 2021 with NICE and EPPI Centre to improve the priority screening algorithms within the EPPI Reviewer software has been expanded to include experts from other European institutions. This collaborative study ($k > 150$ projects) is the largest simulation study of ML approaches with screening, and results will be used to suggest stopping criteria for screening, or when researchers can stop manual screening, as well as provide understandable metrics for researchers to evaluate algorithmic performance. Our role, and NICE's role, is to provide user input regarding the metrics and output of ML-assisted screening. Status: two algorithms have so far been analysed, with eight more remaining. We will provide EPPI with more datasets if needed for training and testing of the algorithms. Next steps are to maintain current collaboration, particularly with EPPI centre and in relation to the priority screening project and user test the stopping criteria once it is drafted.

INSIA Methods Working Group – AI Sub-group

NIPH is one of the member organizations in the International Network for Social Intervention Assessment (INSIA). In May, one of our team members (HB) was elected lead of an AI methods sub-group, which is currently made up of evidence synthesis professionals from the Swedish Agency for Health Technology Assessment and Assessment of Social Services (SBU), French National Authority for Health (HAS) and the Canadian National Institute of Excellence in Health and Social Services (INESSS). The group is working on a strategy for future projects, which will likely entail collaborations across the institutes on how to use AI in social intervention assessments.

International Collaboration for the Automation of Systematic Reviews (ICASR)

ICASR was launched with the aim of seamlessly integrating all components involved in automating the production of systematic reviews. The collaboration, including members from NIEHS (NIH), Cochrane, UCL, CREBP (Bond University), and others, focuses on principles such as efficiency improvement, automation across SR tasks, adherence to high standards, collaboration, open-source practices, and replicability. The first meeting in Vienna (October 2015) outlined these principles (3), laying the foundation for advancing automation in SR production.

The ML team has been invited to sit in a strategic planning group due to our knowledge and experience with implementation and have participated in to two meetings so far. We will continue our involvement to stay abreast of automation developments within the systematic review field and guide the direction for future meetings and conferences. Tasks might involve assisting with organizing and planning the future direction of ICASR and applying for EU funding for networking activities.

Julius Kühn-Institut (JKI)

The ML team has established collaborations with JKI to share knowledge, resources and identify synergies. JKI has created their own systematic review software (CADIMA) and hired an AI researcher to further develop advanced, but user-friendly techniques, whereas NIPH relies on off-the-shelf products. We are both working towards the same goal, but from quite different points of departure, and with different restraints and opportunities. JKI is continuously improving their software, and NIPH has provided them with data that is used as basis for development of semi-automated screening on both T/A and at full text level and using different classifiers for classification of references. They are also exploring possibilities for semi- automation of data extraction, and the ML team have provided input on our wishes for a data extraction function/tool to align with HTV needs for data extraction in our products, as well as providing data for them to evaluate their developments on.

Robert Koch Institute

The Robert Koch Institute reached out to the ML team in May. They are the national public health institute in Germany, and they are exploring the possibilities of identifying existing ML tools for evidence synthesis and creating a workflow as part of a project they are starting on Public Health Impact Analyses. Their aim is to replicate a review that has already been done, with the help of the application of the tools, and to compare the results. In searching for institutes that have made similar efforts before, they came across the work conducted by the ML team and were very interested in exchanging ideas with us and in presenting their project. Currently we have had two meetings where they have shared their work and we have provided them with our experiences as well as input.

The Danish Center for Social Science Research (VIVE)

At the INSIA conference in May, one of the participating ML team members encountered a representative from VIVE, The Danish Center for Social Science Research. VIVE has a collaboration with Campbell, VIVE Campbell, which conducts systematic reviews and other

high-quality reviews in the social sciences domain relevant to the Danish welfare system. The VIVE group have used some of the ML teams' previous reports to inform their ML efforts within the evidence synthesis process. They have amongst other things an ongoing project where they are developing methods and programs for abstract screening using ChatGPT and intend to use the method and program in conjunction with the priority screening function in EPPI Reviewer for a complex screening process in an ongoing review project. The current aims for our collaboration are to exchange knowledge and experiences and to learn from each other, regarding ongoing projects and potential future project collaborations.

Final reflections

Challenges the team has faced

The limited protected time outside of the external funding has posed difficulties, hindering the team's ability to fully engage with ML advancements. Keeping up with the rapid developments in the AI/ML domain has proven challenging due to capacity constraints and we feel we have not been able to keep abreast of important developments in the field.

Ensuring a comprehensive understanding of ML usage across all teams and identifying areas requiring optimization and further learning has also been de-prioritised due to time restrictions. The absence of a rotating membership structure has been felt as a limitation, as time constraints have impeded the inclusion of additional members in the team this year due to the amount of time needed to onboard new members unfamiliar with ML.

The presence of uncertainties around the team's role and objectives has been a recurring issue, particularly in the initial months of the year, negatively impacting both production and motivation. This underscores the importance of clear expectations and guidance from leadership forming a clear mandate for the team’s work, even in challenging times involving reorganizational processes.

What has worked well?

On the other side, ML 3.0 has achieved a great deal during its iteration and there are some key aspects worth highlighting. The interdisciplinary nature of the team has been a key success factor, with the inclusion of a librarian proving particularly valuable. Consistent weekly meetings have played a pivotal role in maintaining team cohesion, preventing potential drift. Furthermore, the team's intrinsic motivation for ML has been a driving force, propelling the group forward.

External funding has been a notable strength, providing crucial support for the team's initiatives. The collaboration with EPPI centre to address technical aspects of the data management software we use has been instrumental, offering valuable insights into implementation and a deeper technical understanding of the ML tools available. The external funding allowed team members to attend conferences which not only facilitated exposure to diverse perspectives but has also fostered networking opportunities and enabling the team to

stay abreast of advancements in the field. We also found that presentations at conferences have been a major contributor to getting our work known outside of the institute and this has generated a lot of interest internationally for our work.

References

1. Muller AE, Berg RC, Meneses-Echavez JF, Ames HMR, Borge TC, Jardim PSJ, et al. The effect of machine learning tools for evidence synthesis on resource use and time-to-completion: protocol for a retrospective pilot study. Systematic Reviews 2023;12(1):7. DOI: 10.1186/s13643-023-02171-y
2. Rose C, Ringsten M, Bidonde J, Glanville J, Berg R, Cooper C, et al. Using a large language model (ChatGPT) to assess risk of bias in randomized controlled trials of medical interventions: protocol for a pilot study of interrater agreement with human reviewers 2023.
3. Beller E, Clark J, Tsafnat G, Adams C, Diehl H, Lund H, et al. Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). Systematic Reviews 2018;7(1):77. DOI: 10.1186/s13643-018-0740-7

NIPH
Norwegian Institute of Public Health

Published by the Norwegian Institute of Public Health
March 2024
PO Box 222 Skøyen
N-0213 Oslo
Tel.: (+47) 21 07 70 00

The report can be downloaded as a pdf at www.fhi.no