NIPH
Norwegian Institute of Public Health

REPORT

2023

Implementation of machine learning in evidence syntheses in the Cluster for Reviews and Health Technology Assessments: Final report 2021-2022

Utgitt av Folkehelseinstituttet
Område for helsetjenester
Tittel Implementering av maskinlæring i kunnskapsoppsummeringer i klynge for vurdering av tiltak: Sluttrapport 2021-2022
English title Implementation of machine learning in evidence syntheses in the Cluster for Reviews and Health Technology Assessments: Final report 2021-2022
Ansvarlig Camilla Stoltenberg, direktør
Forfattere Tiril C. Borge, Folkehelseinstituttet
Heather Ames, med-leder, Folkehelseinstituttet
Patricia Jacobsen Jardim, Folkehelseinstituttet
Jose F Meneses-Echavez, Folkehelseinstituttet
Jan Himmels, Folkehelseinstituttet
Christopher Rose, Folkehelseinstituttet
Christine Hestevik, Folkehelseinstituttet
Ashley Elizabeth Muller, prosjektleder, Folkehelseinstituttet
ISBN 978-82-8406-362-1
Publikasjonstype Report (Rapport)
Antall sider 22
Oppdragsgiver Folkehelseinstituttet
Emneord(MeSH) biomedical; technological assessment, health; unsupervised machine learning; supervised machine learning; deep learning
Sitering Borge, TC, Ames H, Jardim PJ, Meneses-Echavez, JF, Himmels J, Rose C, Hestevik, C, Muller AE. Implementation of machine learning in evidence syntheses in the Cluster for Reviews and Health Technology Assessments: Final report 2021-2022 [Implementering av maskinlæring i kunnskapsoppsummeringer i klynge for vurdering av tiltak: Sluttrapport 2021-2022] Oslo: Folkehelseinstituttet, 2023

Content

Hovedbudskap

Maskinlæring (ML) er et satsingsområde for klynge for vurdering av tiltak, Område for helsetjenester, FHI. Høsten 2021 overtok ML 2.0 arbeidet etter ML 1.0. Denne rapporten beskriver ML 2.0 teamet sitt arbeid, resultater og erfaringer. ML team 2.0s nøkkelprestasjoner inkluderer:

- gjennomføring av en intens maskinlæringsuke med mål om kompetanseheving i hele klyngen,
- oppstart av en studie som vil anslå ressursbesparelsene ved bruk av ML i kunnskapsoppsummeringer,
- formidling av vårt ML arbeid i internasjonale fora, som har tydeliggjort det unike arbeidet ML laget har bidratt med inn i kunnskapsoppsummeringsarbeidet i klyngen,
- støtte innovative bruksområder for ML innen metodevurderinger og oppdateringer,
- utarbeidelse av to finansieringssøknader,
- bidrag inn i andre nasjonale og europeiske finansieringssøknader.

ML-teamet har bidratt til en gradvis tilpassing av klyngens metoder til mer effektive arbeidsflytprosesser, noe som nå merkes i ressursbesparelser i klyngens prosjekter. Ressursbesparelsene gjorde at vi kunne levere flere rapporter enn mulig ved bruk av kun tradisjonelle metoder eller bruke mer tid på andre deler av prosessen.

ML 2.0 fremstår fortsatt som et unikt og innovativt team som leder og tilrettelegger for implementering av ML innenfor kunnskapsoppsummering. Nye ML-aktiviteter knyttet til utforskning og evaluering av nye funksjoner, verktøy eller emner vil gi et mer åpent og flytende ML-miljø der enhver ansatt kan føle mestring og eierskap over ML-funksjoner eller -verktøy. Disse faktorene er avgjørende for Folkehelseinstituttets evne til å tilpasse seg og fortsette å utmerke seg i kunnskapsoppsummeringsfeltet.

Tittel:
Implementering av maskinlæring i kunnskapsoppsummeringer i klynge for vurdering av tiltak: Sluttrapport 2021-2022

Hvem står bak denne publikasjonen?
Folkehelseinstituttet

Key Messages

Machine learning (ML) is a focus area for the Cluster for Reviews and Health Technology Assessments (HTV), Division for Health Services, NIPH. In autumn 2021, ML 2.0 continued the work after ML 1.0. This report describes the ML 2.0 team's work, results and experiences. ML team 2.0's key achievements include:

- implementation of an intense machine learning week with the aim of increasing capacity in HTV,
- initiation of a study that will estimate the resource savings by ML use within reviews,
- dissemination of our ML work in international fora, which has highlighted the unique work the ML team has contributed with within evidence synthesis in HTV,
- supporting innovative uses of ML within health technology assessments and living evidence and gap maps,
- preparation of two funding applications,
- contributions to other national and European funding applications.

The ML team has contributed to a gradual adaptation of HTVs methods towards more efficient workflow processes, which is now reflected in resource savings in HTVs projects. The resource- and workload savings we have experienced allowed us to deliver more reports than possible with using only traditional methods or spend more time on other parts of the evidence synthesis process.

ML 2.0 still appears as a unique and innovative team that leads and facilitates the implementation of ML within evidence syntheses. New ML activities related to exploration and evaluation of new functions, tools or topics would allow for a more open and fluid ML environment where any employee can feel mastery and ownership over ML functions or tools. These factors are crucial for NIPH's ability to adapt and to continue to excel in the rapidly developing evidence synthesis field.

Title:
Implementation of machine learning in evidence syntheses in the Cluster for Reviews and Health Technology Assessments: Final report 2021-2022

Publisher:
The Norwegian Institute of Public Health

Preface

The Cluster for Reviews and Health Technology Assessments, Division for Health Services at the Norwegian Institute of Public Health (NIPH) decided in the fall of 2020 to conduct a project on machine learning related to the conduct of evidence syntheses. The goals were to test and document pros and cons of using machine learning in various phases of the conduct of evidence syntheses, as well as build employees' competence in using machine learning. Based on the work of this team, a new team was established that continued and built on the first team's achievements and challenges; ML team 2.0. The overall goal of the ML team 2.0 was to contribute to ML being used in most of HTV's evidence synthesis products, identify and evaluate new ML functions, as well as further implementation and capacity building activities within HTV.
A team of eleven, consisting of both core and rolling members, worked toward these goals from August 2021 until November 2022. This report describes their work.

The report is relevant for researchers and managers interested in implementing machine learning in their evidence syntheses. It is particularly relevant for evidence synthesis environments that do not have machine learning specialists.

Financing
The work was self-initiated and financed by the Cluster for Reviews and Health Technology Assessments, Division for Health Services at the Norwegian Institute of Public Health

Team members
Ashley Elizabeth Muller, project leader; Heather Ames, co-leader; Tiril C. Borge; Patricia Jacobsen Jardim; Jose F Meneses-Echavez; Jan Himmels; Christopher Rose; Christine Hestevik; Hans Bugge Bergsund; Line Evensen; Severin Zinöcker.

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

Kåre Birger Hagen	Rigmor C Berg	Ashley E. Muller
Research director	*Department director*	*Project leader*

Background

Since early 2020, the Cluster for Reviews and Health Technology Assessments, Division for Health Services at the Norwegian Institute of Public Health (NIPH), became increasingly aware of the potential benefits of using machine learning (ML) in the conduct of evidence syntheses. Thus, the leaders in the cluster decided to initiate a project on ML. Since late 2020, the Cluster for Reviews and Health Technology Assessments has funded a machine learning (ML) team.

The ML team's work was anchored in the preliminary NIPH strategies for the 2019-2024 period concerning automation, increasing speed of evidence syntheses, and workflow and methods innovation. One of the goals of the division-specific strategies was for the Division for Health Services to have an active role in automation and digitalization of work processes, and to use these practices to summarize evidence more efficiently.

Since the team's creation in late 2020, we have been working towards this goal. NIPH has become a leader in integrating ML into evidence synthesis. This team can be seen as a strategic innovation, an attempt to change the "business model" of the cluster to ensure a sustainable competitive advantage over other evidence synthesis providers or environments. ML allows the most effective use of scarce, valuable human resources. Even "fully automated" processes require humans, but at different points - training, interpretation, quality check. ML is meant to do *complex, repetitive* tasks for us, so that we can do other things.

Team 2.0 began informally after the summer of 2021. One team member had left the institution, one new core team member and two new rolling members joined, and the team had delivered both a final report for Team 1.0 as well as strategy report for team 2.0. A new project announcement was published in November 2021, and Team 2.0 was officially financed from that point. In team 2.0 a co-leader with responsibility for implementation and teaching was also added.

Goals

The overall goal of the ML team is to use ML in a way that best combines human intelligence and machine learning, to enhance human activities, by figuring out how best to integrate ML and workflow changes, throughout the review process. The ML team 2.0

have done this by focusing on implementation, innovation and evaluation using iterative and agile methodology.

Team 2.0 had three specific subgoals set out in the team announcement (*lagutlysning)*:

1. to contribute to ML being used in the majority of reviews;
2. to facilitate all review teams having the knowledge and confidence to use at least one ML function by June 2022;
3. to continue to identify and evaluate ML innovations and assess how they can improve workflows and products.

Project results

The following text details team activities undertaken from November 2021. Performance measurement and activities are mapped to the November 2021 project description. This section reports on all activities and their quantitative and qualitative results. Where results have yet been presented or made available, we present them in the text.

Assessment of achievement of team goals

Below we present activities specifically related to each goal, as reported in the team announcement (lagutlysning).

Goal 1: To contribute to ML being used in the majority of reviews

- Using extracted data from the "ML versus no-ML retrospective study", 66% of reviews with a protocol or report published during Team 2.0 (September 2021 - October 2022) have used ML (23 out of 35), and 34% have not (12 out of 35). This sample is limited to those for which KL has resource data and excludes *notater* and other reviews produced for internal commissioners.
- Due to maturity of HTVs use of ML, the ML team is no longer involved in all projects that use ML, only those that request help or those that KL prioritizes to us to receive help. At any given time point, we are not able to count how many projects are or are not using ML. Since we began formally tracking help requests in October 2021, 26 teams or projects have received or are currently receiving help.
- An important development is the increasing involvement of the ML team in projects related to Nye Metoder and different types of health technology assessments.

Goal 2: To facilitate all review teams having the knowledge and confidence to use at least one ML function by June 2022

- In the ML Week 2022 evaluation, 60% of particpiants reported feeeling comfortable enough to use ML in their next project independently (9 of 15).
- We estimate that all teams are able to independently use at least one ML function. This is because there are no longer teams requesting help where all members are ML naïve, as at least one team member are i) familiar with and comfortable using at least one ML function and ii) they can to a certain extent independently reflect

around how ML can be used in a project. there are no longer teams requesting help that report that all team members are ML-naïve.

- We are helping teams that contain one or two new or less experienced employees, or teams contributing to an evaluation.
- Another indication of knowledge and trust is reported intention to use ML in future projects. After the September 2021 cluster seminar, 16/17 employees who filled out the evaluation form reported they intended on using ML in their next project, and 1/17 reported "no". After the March 2022 "ML Week", 14/16 employees reported planning on using ML, 1/16 reported "no", and 1/16 was unsure. Because the latter evaluation was anonymous, we are not able to track whether ML Week evaluators were the same sample, or if, as we hope, they were primarily a new batch of employees who did not attend the 2021 klyngeseminar.
- If this goal will remain a performance indicator, we will need to quantify and monitor developments in knowledge and trust. The HTV ML quizzes are one way forward in the short term. We have also planned distinct work packages/activities in both the NFR *Human-machine teaming* application and the DFØ *Kompetanseheving* application to regularly monitor and improve knowledge and trust. Per end of January 2023, the DFØ application has been granted and the NFR application is pending.

Goal 3: To continue to identify and evaluate ML innovations and assess how they can improve workflows and products.

- Exploration and use of OpenAlex in the production of living reviews in different areas, such as COVID-19 (Omicron living map) and review updates.
- Exploration and use of ChatGPT to draft the Key Messages of the ML strategy report automatically, by copying the text of the suggested strategy section and requesting a summarized version. We then requested a Norwegian translation of the summary, which became the *Hovedbudskap*.

For additional activities, we refer the reader to the sections below on Evaluations and Innovation.

Time and resources

Team 2.0 was allocated a maximum of 1,4 full-time equivalents. Figure 1 illustrates team turnover according to rolling member status and periods of leave, beginning just before the *lagutlysning* was sent out. In this team, the lead (AEM) was responsible for the team and innovation and evaluation projects. The team co-lead (HMRA) had responsibility for implementation and teaching projects. The co-lead took over to lead the team from August-October. When the lead returned in November the co-lead left the team.

Additionally, the team had four core-members; these members used between 20-40 % of their time working with machine learning during their period of participation in ML 2.0. Moreover, the team had four rolling members, these members had a smaller percentage of involvement and were mainly involved in specific projects. This was quite useful as there were some projects that needed more people involved or a specific competence that the rolling member provided. The period where the team members were actively involved in ML 2.0 is illustrated in Figure 1.

Finally, the team had an advisor, Chris Rose, who contributed about 5 percent of his time to the team. The advisor had advanced knowledge of ML and contributed to strategy planning, discussions, and questions from the team.

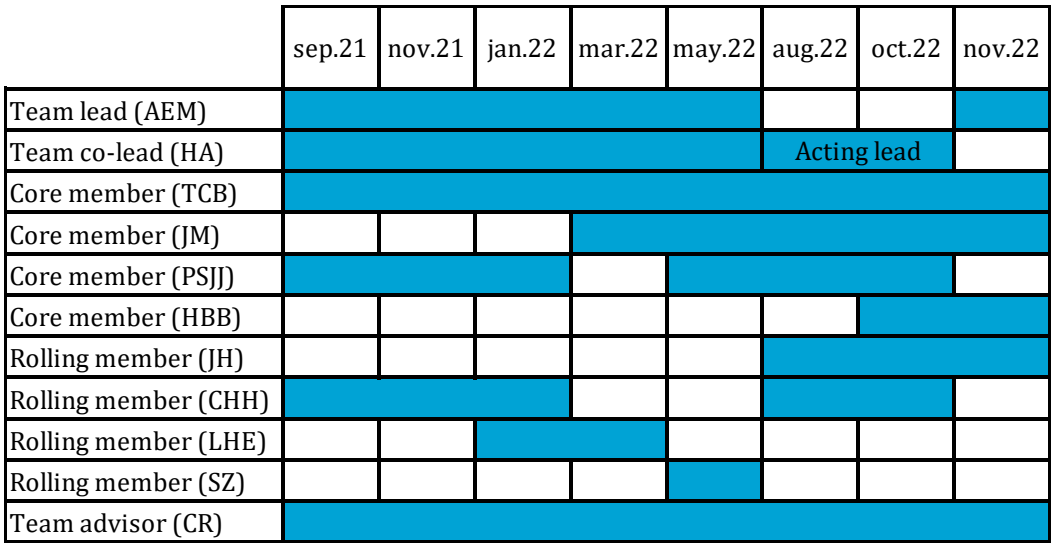

Figure 1 *ML 2.0 team members*

Internal team capacity building

In the ML 1.0 team, we relied on an intensive peer-teaching program. This was repeated during the fall of 2021 but needed significant changes. In Team 2.0, we retooled this peer-teaching program into a distinct syllabus that we then used during onboarding of new members. If this syllabus is updated at least quarterly, and updates can be drawn from the innovation activities, it can be re-used with future team members. An important component of this onboarding syllabus and process is that the new members being onboarded must have a high level of self-organization and drive, while at the same time receiving intensive follow-up and learning monitoring from an existing team member.

Implementation and training

Also included in capacity-building to HTV is the one-on-one support provided by «ML contacts», that is, by ML team members to review-teams who have requested support.

Since we began a formal system of support frequent in the end of October 2021, 26 projects have requested support online, for a variety of different review types.

There has been an overall increase in the heterogeneity of projects asking for and receiving ML help, pointing towards a positive spread of ML throughout HTV's portfolio of products. A recent achievement is the use of ML within health technology assessments, single technology assessments and related products, as these were conspicuously absent in team 1.0. Several help requests have been logged by KL, demonstrating the importance of vocal KL support and "pushes" towards ML. The team has also been asked to provide guidance on horizon-scanning and *metodevarsling* projects.

As agreed with KL in early 2022, resource use is charged to the review team, rather than to the ML team. We estimate that providing ML support to a team uses on average 8.5 hours per project. This, however, can have a lot of variation based on the team's experience level and the complexity of the ML functions being implemented.

Other implementation and training highlights

- A standard operating procedures guide for the team that describe procedures for the team's activities, quality control mechanisms, guidelines for the development and implementation of new training materials and onboarding processes.
- «ML week»:
  - A one-week learning festival where employees were invited to learn about the conceptual aspects of the ML functions we use and the technical knowledge on how to implement them. All presentations are available on the ML teams SharePoint site
  - Attendance summary (in the ML week to KL presentation)
  - This week provided a clear indication from leadership that ML was an expected part of the review process going forward
  - The teaching materials from this week are easily scalable and can provide the basis for developing future courses
- Quizzes to monitor HTV knowledge needs and growth (underway)
- All review-specific protocol and report templates are updated with ML-language (complete). A new ML appendix was developed in the fall of 2022 to address feedback from end users. The new appendix was user tested and finalized in early 2023.
- Training materials updated to reflect results of further evaluations and learning
- OpenAlex implemented across several reviews and has become widely accepted as the third database for searching
- Successful transition away from one-on-one help for most projects to teams working independently and accessing support materials in the ML SharePoint room

Evaluations

A portion of the ML innovations identified by Team 1.0, were then prioritized for evaluation during Team 2.0, as were newer innovations. Evaluation activities were meant to

provide the evidence base bridging innovation and subsequent implementation in HTV. The following activities are first described according to status of *complete*, *ongoing*, *waiting*, or *tabled*.

Complete

1. We have developed an evaluation protocol that we use to assess the need, potential gain, risk, and administrative requirements of a potential evaluation, and that requires peer-review from the ML team. This process has proved extremely valuable, as it first requires a clear owner (the evaluation lead), that conducts a brief review of the existing evidence base, thus preventing unnecessary duplications as well as clearly assessing when an existing evaluation from another instituion is not sufficient, helps the team think through what "success" would look like and how measureable this is, and provides a transparent and replicable protocol when completed..
2. Based on our evaluations and on recently published studies, we recommend rolling out ***OpenAlex***[1] to replace a portion of traditional academic databases. Our latest evaluation , and the first that was not limited to COVID-19, demonstrated that studies retrieved from OpenAlex were more than three times as likely to be relevant than those identified from traditional searches. This enables review teams to begin screening faster and to identify (or screen) far fewer irrelevant studies.
3. We evaluated ***clustering to confirm irrelevance*** in a small randomized crossover experiment, in which four ML team members were randomized in pairs to help screen for an entirely new review, with inclusion criteria that they had to learn. The two conditions were screening according to manual procedures of their own choosing, or with the ability to use clustering. Main result: clustering in this procedure did not improve the hourly speed of screening, with a mean difference of 194 more studies/hour (from 54 fewer to 442 more), in the clustering arm compared to the non-clustering arm. Based on the limitations of this experiment, particularly that participants were new to the project and reported therefore being more hesitant to bulk-exclude, and based on our previous published study demonstrating clustering's efficacy and precision, we continue to recommend that clustering be used by teams who are already familiar with a project.

[1] OpenAlex is a knowledge graph and an "open source" dataset with more than 250 million scientific objects such as articles, white papers, reports and conference abstracts (Priem, Piwowar, & Orr, 2022). The dataset is composed of five types of scholarly entities (works, authors, venues, institutions, and concepts) and the connections between them. Instead of searching according to words found in a study title or abstract, MeSH terms or keywords provided by the author, journal or database, OpenAlex uses deep learning to link these objects together, in addition to bibliometric and citation similarities.

Ongoing

1. ML versus no-ML retrospective study. This ongoing study is the first to our knowledge that will quantify the effect of ML on resource use and time-to-completion, on an organizational level. We have also operationalized over-use and under-use of ML, which will provide NIPH and other organizations with guidelines for future implementation and quality control. This study has strengthened collaboration with King's College London via co-authorship with Chris Cooper, Associate Director for Service Transformation. Several international research environments have expressed their interest in our work and willingness to collaborate in a future, prospective study. The study protocol was accepted for publication in January 2023 in BMC Systematic Reviews, and is available online
2. An ongoing project from ML 1.0 is the collaborative ***priority screening*** algorithm improvement project. Preliminary results suggest we will not change our current recommendation of using this function with EPPI Reviewer. Ideally we will end up with statistical stopping criteria, plus an improved function in EPPI Reviewer that assists researchers in deciding when they will change screening practices.
3. Retrospective evaluation of the utility of custom classifiers to assist in screening and potentially sampling for ***qualitative evidence syntheses***.

Waiting or tabled

1. Waiting: An evaluation protocol related to using clustering to populate an evidence and gap map is ready to be used, once an appropriate commission is identified
2. Tabled: An evaluation of the free software ***Rayyan's*** ranking algorithm function compared to EPPI Reviewer's priority screening function.

Main challenges to evaluation activities

Quantifying workload savings is difficult when project members are not used to tracking and reporting time for specific tasks. The best practice is when the employee leading the evaluation is also the ML contact, or otherwise embedded in the project team.

Next steps

- Strengthen the use of the evaluation protocol – this was a popular topic at the International Collaboration for the Automation of Systematic Reviews' annual meeting in June 2022.
- Team 3.0 proceeds with scalable training material for OpenAlex.

Innovation

Innovation activities refer to the team looking outwards to identify new ML functions, new applications of existing functions, or other novel ways of using ML to improve our products or workflows. A selection of innovation activities is prioritized for evaluation

and have been described in the previous section. The activities in this section are those that have not (yet) been evaluated.

- ***Automatic review updates.*** Another application of OpenAlex is to allow it to harvest new, relevant studies, using an older review's included studies as seed studies – rather than re-searching academic databases. Extensive evaluation of this purpose has been conducted using COVID-19 reviews both externally, during Team 1.0, and most recently during our third update of a rapid review regarding children. As agreed upon with KL, we are proceeding with scalable capacity-building materials, rather than planning another evaluation.
- ***Automatic assessment of documentation package in a single technology assessment (STA)***. Two STAs have experimented progressively with using OpenAlex to quality-control documentation packages. Experiences are being gained and an evaluation protocol written, but documentation has yet occurred.
- ***Semi-automation of a living evidence and gap map*** using a pipeline of neural network automated study retrieval and a suite of custom classifiers to categorize studies. The omicron living map (Sasha Poulsson / Kjetil Bruberg) has been a test case of increasingly advanced ML function use, and embracing living review products to rapidly respond to commissioner needs. This map was created to deliver new studies to *Smittevern*, and has demonstrated – although not fully documented – the use of these functions.
- ***Combination of ML and agile methods:*** Demonstrated feasibility of combining ML with agile project methods in a review on overdose warning systems (link). We were able to complete the review in only 180 hours attributed mainly to the innovative use of ML combined with agile methodology.

Main challenges to innovation activities

- The fields of ML and AI continue to move at lighting speed. During a six-month span in which we put together "ML Week", prepared for and delivered a suite of presentations and workshops at ICASR and IQWiG, and then took July holidays, nearly 200 relevant studies were published. Keeping abreast of innovations in the field requires teammates with the motivation, capability, and time set aside to simply explore, as well as maintaining contact with our networks.

Next steps

- The use of OpenAlex and custom classifiers to regularly update and populate a living evidence map does not need to be evaluated in-house, as the EPPI Centre environment has recently published a cost-effectiveness and performance study in the form of an eight-arm RCT, which document that OpenAlex is more precise and more cost-effective than traditional databases for covid-19-related reviews.
- Finish training materials for using custom classifiers to categorize studies on the title/abstract level, create training materials tailored to updating reviews, and support further implementation.

Dissemination and collaboration outside of the ML team

The ML team remains a uniquely innovative team in the evidence synthesis world. A large part of this achievement is due to intentionally surrounding ourselves with experts – individuals and groups from whom we can learn. Disseminating our achievements, results, lessons learned, challenges, and plans has been a successful way to demonstrate our expertise, to signal our interest in collaboration with external experts (referring to people external to the field of reviewing and/or to our organization), and to open pathways for our own learning.

Dissemination and collaboration activities

The following tables display our structured dissemination and collaboration activities.

Table 1: *Overview of dissemination and impact*

Dissemination		Novelty and impact
31.10.22	Ukestart: Omikron litteratursøk – til Omikron kart	Sasha Poulsson presented the process and results of her innovative use of classifiers and OpenAlex to reduce manual time needed to update a living map.
19.11.22	Presentation at XVIII Conference of the Iberoamerican Cochrane Centre. Barcelona, Spain. Title: Machine learning to accelerate evidence production: experience from the Norwegian Institute of Public Health	Dissemination of our agenda in innovation and evaluation to researchers and stakeholders within evidence synthesis. Positioning NIPH as a leader in ML use for evidence synthesis.
11.10.22	Presentation at What Works Global Submit (WWGS). Implementation and evaluation activities to build support for machine learning in a systematic review organization	Connected with the lead of a developing ML/reviewing environment at Newcastle University, who asked for team lead mentoring.
11.10.22	Presentation at WWGS. How much time can we save screening in a systematic review by using machine learning functions?	Dissemination of our agenda in innovation and evaluation to researchers and stakeholders within evidence synthesis and policy making. Positioning NIPH as a leader in ML use for evidence synthesis.
10.6.22	Keynote speech at Information Retrieval Meeting (IRM) in Köln, Germany, led	Contributing to establishing NIPH as implementation leader within the field of ML in evidence synthesis. Connected with Bond

	by the Institute for Quality and Efficiency in Health Care. Machine learning in evidence synthesis: who are the players needed for implementation?	University (co-keynote holder), which became a collaboration partner on the NFR grant, held two automation workshops for librarians, and a co-author of the "agile reviewing" article.
10.6.22	Workshop at IRM: Implementation and evaluation activities to build support for machine learning.	Workshops focused on different stages of integrating ML into NIPH, from onboarding new team members and keeping up to date and creating training materials and train the trainer, to embedding process and performance evaluations into existing commissioned reviews
11.6.22	Joint workshop with Julius Kühn-Institute, another innovative evidence synthesis environment at IRM: Two roadmaps for using machine learning in evidence synthesis, across disciplines	Provided information on the step-by-step process of how ML has been successfully implemented at NIPH, including suggestions and resources, providing a detailed road map of how ML can be implemented at other institutions enabling other institutions. Further established NIPH as implementation leader within the field of ML in evidence synthesis.
11.6.22	Presentation at IRM: Is RobotReviewer, a semiautomated risk of bias tool, acceptable to researchers?	Presented novel qualitative results from our mixed methods randomized trial.
11.6.22	Presentation at IRM: A digitalization project case study: designing a cross-software solution to standardize, share, and re-use systematic review data	Presentation of the previously planned project to purchase a software solution to facilitate storage, sharing, and re-use of data during the review process.
11.6.22	Presentation at IRM: When can we stop screening studies? A cross-institutional simulation study	Presentation of preliminary findings of a multi-institutional study to develop statistical stopping criteria for screening, by James Thomas from UCL/EPPI Centre, where NIPH is collaborator.
8.6.22	Presentation at International Collaboration for the Automation of Systematic Reviews, 7th network meeting: Other institutions present: UCL, Bond, Cochrane Netherlands,	Presentation of ML team's evaluation activities and recruitment to a planned prospective study. Very important networking session, as we were able to establish ourselves as leading implementors

	EFSA, EvidencePrime, Scion, Evidentlia, Epistemonikos	
24.1.22	Ukestart. Automatic text clustering vs. human categorization	Connected with Kim Kristoffer Dysthe, who peer reviewed NFR grant and is now conducting a research exchange at Bond University.
Ongoing	Contribution to a Campbell Collaboration handbook chapter about screening and sampling for qualitative evidence syntheses	Upon publication, this will be the first guidance we are aware of around ML in qualitative reviews.
Tabled	Proposal for HTAi International meeting in 2023. How is artificial intelligence impacting the field of HTAs? How do we build the skills and knowledge we need?	Developed a text and structure for an expert panel as well as a pool of potential experts

Table 2: *Manuscripts and pre-prints*

Citation	Novelty and impact
Jardim PSJ, Rose CJ, Ames HM, Meneses-Echavez JF, Van de Velde S, Muller AE. Automating risk of bias assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a machine learning system. *BMC Med Res Methodol*. 2022 Jun 8;22(1):167. https://doi.org/10.1186/s12874-022-01649-y	The only randomized study and the only mixed methods study of the quantitative performance and researcher acceptability of a ML system to assess risk of bias.
Muller, A. E., H. M. R. Ames, P. S. J. Jardim and C. J. Rose (2022). "Machine learning in systematic reviews: Comparing automated text clustering with Lingo3G and human researcher categorization in a rapid review." *Res Synth Methods* 13(2): 229-241. https://doi.org/10.1002/jrsm.1541	We believe this is the first study in this area. Systematic reviewers without machine learning expertise can successfully implement automated text clustering. Automated text clustering can provide useable and valid categorizations of text. The time saved compared to human categorization outweighs the time needed to sort through and make sense of the automated categories.
Borge, T. and A. Muller (in press). " Overdosevarslingssystemer – en kartleggingsoversikt med maskinlæring." *Nordisk alkohol- & narkotikatidsskrift*.	The first paper to our knowledge, published in a Scandinavian-language journal, that has used a suite of ML functions, in-

	cluding a neural network search to supplement academic databases and the automatic screening of 75% of retrieved references. Contributes to dissemination of advantages of integrating ML in workflow changes in the Nordic research community.
Muller AEM, Berg RC, Clark J, Cooper C, Kornør H, Borge TC (in progress). Agile systematic reviewing: a proof of concept.	We introduce a new, data-driven approach to facilitate rapid production of systematic reviews: "agile systematic reviewing", combining a customer-value focus with full integration of ML. This approach can be a tool for the larger evidence synthesis community. Internally, it may help us consistently reduce production resources.
Muller, A., H. M. R. Ames, T. Borge, C. Hestevik, J. F. Meneses-Echavez JF, and C. J. Rose. "A protocol to evaluate unsupervised text clustering to screen and categorize studies in systematic reviews." Posted as a preprint: https://www.researchsquare.com/article/rs-1644531/v1	This study protocol establishes the ML team at NIPH as a leader in the innovative evidence-based application of ML to various review products. It is available for comments and feedback online.
Muller, A.E., Berg, R.C., Meneses-Echavez, J.F. et al. The effect of machine learning tools for evidence synthesis on resource use and time-to-completion: protocol for a retrospective pilot study. Syst Rev 12, 7 (2023). https://doi.org/10.1186/s13643-023-02171-y	This ongoing study bolsters the uniqueness of our team in conducting the first evaluation of the effect of ML adoption on resource use and time-to-completion at organizational level. See the innovation spotlight in *Goal 3: Innovation and evaluation.*

Dissemination channels

We are continuously posting on all our activities and projects at ML team's sites. This helps us disseminate our activities worldwide, reaching wider audiences:

- Open Science Framework: The ML team's site in OSF acts as a living repository of our protocols as well as supplemmentary information (e.g., our syllabus being presented at IRM 2022, Cologne). Documents archived here can be directly cited in institutional report or other publication formats.
- ResearchGate: in addition to what is described above for OSF, our ResearchGate portal allow us to connect with researchers and organizations, and to disseminate our activities (e.g., ongoing projects, reports, and scientific publications).

Moreover, these dissemination channels boost ML team's identity and recognition by target audiences, such as other national institutes of health, research centers, and academia/academics.

Collaborative partners

Below we highlight a couple of central collaborative institutions. For further partnerships, see Funding applications and research projects.

Julius Kühn-Institut (JKI)

We have established collaborations with JKI to share knowledge, resources and identify synergies. In NIPH ML was initially, and sometimes still, seen as disruptive to methodological gold standard approaches. In agricultural science (JKI), systematic review methods are only recently scaling up. 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 very 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. 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.

In June 2022 we conducted a joint workshop with JKI at IRM, with the overarching aim of bringing people together and facilitate future cooperation. During the workshop we first compared our organizations and disciplines' approaches to reviews and ML. Then JKIs programmer provided an introduction of the basics of ML within reviews for non-specialists. Then participants selected facilitated small groups to join, based on topics they wished to brainstorm with others, where topics were focused on implemented strategies for how the ML team has successfully introduced ML at NIPH (Teaching and training, embedding evaluations into commissioned products, and onboarding of new members and keeping up to date), as well as how can we build reviewer trust in ML. The workshop was a success – many participants joined in, good discussions were made, and participants were very enthusiastic. We received great feedback from participants afterwards.

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

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: analyses are in their final stages. EPPI Centre will possibly add more datasets to create a larger data base to train the algorithm on. Next steps are to maintain current collaboration, particularly in relation to the priority screening project, as well as contribute to custom classifier documentation needs.

Networking at NIPH

We put together a cross-division pool of experts and called this the "ML/AI Network at NIPH". We held four formal meetings between March and August 2022: a kick-off meeting, a discussion on bias and reflexivity in the context of machine learning, a discussion around scalability, and a specific meeting to gain feedback on the human-machine teaming grant being written. In these meetings, we actively sought out alternative ways to problem-solve.

Network members provided the team with ad hoc mentoring around implementation, scaling up, and change management, provided feedback on the NFR grant, and in turn requested team information for different NFR grants. In addition, we used the network to promote new publications of network members and to disseminate national and international resources.

Status: This network has been informally organized, without funding or mandate. Anchoring it in HTV or the division, or connecting it to the *Forsknings- og innovasjonsutvalget*, is necessary for it to continue sustainably.

Funding applications and research projects

One major activity that was not part of the team's original mandate was applying for funding. The following funding applications were sent:

- **"Human-machine teaming"** Innovation project submitted to the Norwegian Research Council, Sept 2022. Partners: SINTEF, Bond University. 14 million NOK
- **"Smartere arbeid, ikke hardere arbeid, med maskinlæring i Folkehelseinstituttet"** Capacity-building fund application submtited to the Norwegian Government Agency for Financial Management, Nov 2022. Internal collaboration with *tillitsvalgte*. Applied for 2,43 million NOK (Received 1,46 million NOK).
- **"A Criticality Assessment Framework for Real-World Evidence of Human-centered AI-based Algorithms in Medical Diagnostics"** Letter of intent to be a partner in an EU doctoral network connected to the existing, funded "eBrains" project. Partners: University of Oslo, DNV, Oslo University Hospital, Aix-Marseille University, Charite Berlin, De Montrofrt University, Universidad de Granada, eBrains, Norwegian Artificial Intelligence Research Consortium, among others
- Submitted interest as a partner institution for two types of Horizon Europe 2023-2024 grants.
- Application planned for Stimulab

Conclusion

The ML team 2.0's major achievements included running an intensive, cluster-wide introductory learning week, beginning a study that will estimate the resource savings of ML use within reviews, supporting increasingly innovative uses of ML within health technology assessments and living evidence and gap maps, submitting two funding applications, and contributing to other national and European funding applications.

Together with support from the Cluster management, the ML team has successfully and continuously adapted our methods towards more efficient workflows, which are being noticed in the workload savings that most of our project teams have expressed. This gives the cluster the ability to shape the rapidly changing and fast-pacing environment of evidence-informed decision making. Proof of this innovation is the capital role of ML in the number of updated reviews and living- and semi-automated evidence maps the cluster has published recently. We anticipate that the ML team represents a cornerstone within the transformation of our cluster into a more innovative organization. The resource- and workload savings we have experienced allowed us to deliver more reports than possible with traditional methods.

The ML team remains a uniquely innovative team in the evidence synthesis world and have firmly established us as an implementation lead in the field. Further, ML activities pertaining to exploration and evaluation of new functions together with the freedom to explore and evaluate new topics or functions would allow for a more open and fluid ML environment where any employee could feel mastery and ownership over ML functions or programs. These factors are crucial for NIPH's ability to adapt and to continue to excel in the rapidly developing evidence synthesis field.

Published by the Norwegian Institute of Public Health
February 2023
P.O.B 4404 Nydalen
NO-0403 Oslo
Phone: + 47-21 07 70 00
The report can be downloaded as pdf
at www.fhi.no/en/publ/